Genomic Analysis of Viral Outbreaks

Genomic analysis of viral outbreaks

a dissertation presented
by
Shirlee Wohl
to
The Committee on Higher Degrees in Systems Biology

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Systems Biology

Harvard University
Cambridge, Massachusetts
April 2018

Dissertation Advisor: Professor Pardis C. Sabeti Shirlee Wohl

Genomic analysis of viral outbreaks

Abstract

Due to recent advances in sequencing technologies, genomics has emerged as a powerful tool for combating viral outbreaks. In this dissertation, I show how genomic data can be used to understand the transmission, as well as the origin and evolution, of three different viruses. In Chapter 2, I describe Ebola virus (EBOV) sequences from the 2014–2016 outbreak in Sierra Leone, and what they tell us about cross-border and individual transmission, as well as evolution of the virus. In Chapter 3, we apply these findings to the 2014 EBOV outbreak in Nigeria. I present a direct comparison between genomic data and contact tracing performed during the outbreak, and show that these methods suggest similar transmission patterns, but that viral sequences provide additional information about the origins of the outbreak. In Chapter 4, I describe our efforts to characterize the movement of Zika virus (ZIKV) throughout the Americas in 2016, and I explain how we dealt with challenges related to low viral content in ZIKV samples. In Chapter 5, I explain how mumps virus (MuV) sequences from 2016–2017 informed our understanding of disease spread at multiple geographic scales. I explain the evidence for ongoing MuV transmission in the United States, and show that pairing genomic and detailed epidemiological data reveals transmission within a local community. Taken together, these chapters demonstrate the emerging capabilities of genomics, including how genomic analysis can inform our understanding of viral transmission, and show how viral sequence data can influence public health at local, national, and international scales.

Contents

Acknowledgments

This dissertation would not have been possible without the support of many people. In addition to the contributions described in each chapter, I would like to briefly thank the following individuals for their contributions and encouragement, either personally or professionally, during my time in graduate school:

My advisor, Pardis Sabeti, for her continued support and sage advice, for the opportunities she gave me, for believing in me even when I doubted myself, and for being a constant inspiration to me and everyone around her;

Past and present members of the Sabeti Lab, who continue to amaze me with their passion, compassion, and genuine desire to do good in the world: Hayden Metsky, for countless scientific discussions, for his impressive attention to detail in every scientific pursuit, and for being the only person who has seen me awake at 4am; Steve Schaffner, for many insightful discussions, for shortening my sentences, and for inserting a little humor into each day; Danny Park, for being an excellent mentor and guiding me through my first few years of graduate school; Bronwyn MacInnis and Kayla Barnes, for being constant sources of good advice, and for listening when I needed it most; Nathan Yozwiak, for letting me knock on his door multiple times a day, and for always offering sound advice; Catherine Freije, for making long days in the lab more enjoyable; Chris Matranga

and Katie Siddle, for teaching me how to prepare samples for sequencing, and for their constant good cheer; Bridget Chak, for helping me navigate review boards and complex collaborations, and for always asking how I'm doing; Aaron Lin, for his endless enthusiasm for biology, and for always being available to help; Anne Piantadosi, for her insightful comments and attention to detail; Liz Brown and Samar Mehta, for being wonderful officemates and conversationalists; Sarah Winnicki, for keeping everything running; Dolo Nosamiefan, for our many afternoon chats; the rest of the Sabeti Lab, for support, camaraderie, and always excellent feedback;

Yonatan Grad, for a wealth of scientific ideas and many helpful conversations, for inspiring me to be a better, more thoughtful scientist, and for his feedback as a member of my dissertation advisory committee; Curtis Huttenhower, for his helpful feedback as one of my committee members, for allowing me to rotate in his lab, and for his willingness to serve on my thesis committee as well; Bill Hanage, for his enthusiasm and thought-provoking ideas as a member of my dissertation committee; Dan Neafsey and Marc Lipsitch, for taking the time to serve on my thesis committee;

Sam Reed and Liz Pomerantz, for making Systems Biology a department I could brag about, and for being unfailingly cheerful and helpful in every situation;

My dear friends, too many to name, who keep me inspired, motivated, and happy: the current and former residents of our delightful Amory Street home, Caroline Jaffe, Aviva Musicus, Jacob Evelyn, and Astrid Pacini, who are always there and always supportive, especially Caroline Jaffe, with whom I have had the great privilege of living with for nearly seven years; Maddy Howe and Marj Berman, for countless hours on the phone, discussing every aspect of our lives; Anna Green, for helping me get through graduate school, for knowing every struggle, yet still inviting me over

to play board games; Charlie Fulco, for always stopping to chat in the hallway; my Siege teammates and the entire Boston Ultimate community for keeping me sane (and fit), especially Lizzie Jones, who shared my dream and helped make it a reality; the ladies of Salon, who always have interesting things to say, and never fail to remind me that I am not alone;

And, of course, my family, for always listening and sharing their love: my parents, Mina Levinsky-Wohl and Peter Wohl, for giving me confidence and advice, for always asking questions, for fostering my love of math and science, and for encouraging me at every step along the way; my sister, Nureen Wohl, for showing me how to be a better person, for always seeing the best in me, and for listening even when I'm not ready to talk; and Tsuki Hoshijima, for being my partner in everything, for always having interesting and insightful comments, for making me calmer, for proofreading this dissertation and way too many emails, and for teaching me to be a warmer and more thoughtful person every day.

Chapter 1

Introduction

Preface

Recent advances in sequencing technologies have allowed genomic analysis to play an important role in understanding outbreaks of viral pathogens. This chapter describes current approaches and methods for sequencing viral pathogens and performing phylogenetic, evolutionary, and transmission analyses. It is, in essence, the blueprint for the analyses performed in all other chapters of this dissertation.

This chapter is based on the following work: Wohl S., Schaffner S.F., Sabeti P.C. *Annual Review of Virology*, 2016 [1]. My advisor, Pardis Sabeti, gave me the opportunity to write a review on viral genomic analysis, and I conducted a literature review on this topic, drawing examples and ideas from work I had already done on the subject, described in Chapters 2 and 3. I then wrote the paper and refined its contents with Steve Schaffner.

Of course, the methods and tools presented in this review are merely a snapshot of a rapidly

changing field. Throughout this chapter, I note additional tools and ideas that have emerged since the original publication of the review in 2016, and highlight more recent applications of these methods (see Sections 1.2.2 and 1.4.2). While techniques for generating and analyzing genomic data continue to evolve, the fundamental questions during disease outbreaks — where did the virus originate, how is it changing, and where will it go next — remain the same. By detailing the ways in which genomic data can help answer these questions, I hope to provide a resource that may assist study of future viral outbreaks.

In subsequent chapters, I demonstrate the use of many of these methods to analyze recent viral outbreaks of three different viruses. In Chapter 2, I describe our investigation of Ebola virus (EBOV) both during and after the 2014–2016 outbreak in Western Africa. Chapter 3 describes our continued work on EBOV, this time focusing on analyzing viral transmission within a small, contained outbreak in Nigeria in 2014. In Chapter 4, I explain how we used genomics to understand the spread of Zika virus (ZIKV) in the Americas in 2015–2016, with an emphasis on how we dealt with sequencing challenges unique to ZIKV. Finally, in Chapter 5 I explain how we used viral sequences to understand mumps virus spread in the United States in 2016–2017. This final chapter is the culmination of many of the lessons learned in the preceding chapters, and emphasizes the added value of genomics in public health investigations and the power of combining genomic and epidemiological data for outbreak response.

1.1 A Brief History of Viral Outbreak Investigation

Human history is replete with viral disease outbreaks that have devastated communities and entire populations. Famous examples include smallpox, the 1918 Spanish flu pandemic, the ongoing HIV/AIDS pandemic, the 2009 H1N1 influenza pandemic, and the 2014–2016 EBOV epidemic in Western Africa. Not all outbreaks reach pandemic status; many are contained by environmental factors or control measures. Regardless of the final number of individuals affected, outbreak investigation always has the same two aims: termination of the current outbreak and prevention of future ones.

For decades, epidemiological methods such as detailed contact tracing and mathematical modeling have been used to support these aims [2–5]. Although those methods have worked well for stemming outbreaks of low-prevalence diseases like severe acute respiratory syndrome (SARS), their effectiveness is limited for large outbreaks, diseases with long latent periods, and outbreaks that occur in remote areas [6]. For these kinds of outbreaks, it is difficult to collect the detailed observations needed to parameterize epidemiological models with predictive power.

Applying molecular biology tools to traditional epidemiology has greatly improved outbreak monitoring and prevention for all types of viral diseases. These tools include genotypic and phenotypic methods to determine the specific strain or type of virus circulating in a population. They can be used to improve diagnostics, to guide treatment programs and vaccine development, and to trace the spread of pathogens [7–10].

Moving to full genomic analysis expands our capacity to understand viral outbreaks even fur-

ther, because nucleotide-level resolution can distinguish isolates of the same viral strain. For example, when the World Health Organization issued a global alert for SARS in 2003, the disease-causing agent was still unknown. Subsequent sequencing of isolates and identification of SARS coronavirus as the pathogen responsible led to development of sequence-based diagnostics necessary for the remarkable containment of the outbreak [2, 11–13]. In 1992, viral sequencing was also used to supplement epidemiological investigation when a patient claimed she had contracted HIV through a dental procedure. Phylogenetic analysis of HIV from the dentist and five of his dental patients — which showed that the viruses were closely related — made it clear that the dentist had indeed transmitted HIV to his patients [14]. In these two examples, whole-genome viral sequencing led to advances in diagnostics and in understanding transmission, respectively, that were not possible using other methods.

The 2014–2016 EBOV epidemic in Western Africa has provided one of the first applications of near-real-time whole-genome viral sequencing to understand a disease outbreak from its onset. Genomic analysis during the outbreak was made possible by recent advances in high-throughput sequencing, computational methods, and data processing. This epidemic also spurred the development of numerous methods for exploiting whole-genome sequencing in future outbreaks.

Here, we compile and describe existing methods for analyzing genomic data from viral disease outbreaks. We focus on fundamental questions and how genomic data can be used to answer them. Although we include examples from a number of viruses in our review, we use the EBOV epidemic as the primary example throughout, both because rich genomic data are available for that outbreak and because many of the analyses described here were applied during it.

1.2 Viral Sequencing Methods

Accurate sequencing is key to producing high-quality genomes for analysis. Dramatic improvements in high-throughput (also known as next-generation) sequencing technologies and in virus-specific sequencing [15–18] in the past decade have enabled sequencing of viruses, known and novel, from all kinds of samples. We briefly review current sequencing technologies and discuss methods for sequence processing.

1.2.1 Current Sequencing Technologies

It is essential that patient samples be processed and sequenced in a way that will provide the highest quality data for downstream analysis. For RNA viruses, timing is especially important: degradation can occur quickly in clinical samples and is common, so the time between sample collection and sequencing should be minimized ([19]; see [20] for procedures used in an EBOV diagnostic laboratory). In general, all sample processing should consider both sample preservation and researcher safety [21].

Sequencing itself has progressed from technologies tailored to a specific viral sequence to sequence-independent, high-throughput approaches. Amplicon-based sequencing is the most common sequence-dependent method and was used early in the EBOV outbreak [22]. In that study, viral genomes were amplified in long (often $\geq$2 kb) overlapping fragments by reverse transcriptase polymerase chain reaction (RT-PCR) with EBOV-specific primers; these fragments were then sequenced by Sanger sequencing. This method is popular for detecting and studying viruses

because it is fast and can be used to amplify very small amounts of material.

The speed and accuracy of amplicon-based sequencing has made it an effective method for on-site sequencing even in remote field settings [23], and the resulting genomes are sufficient for pathogen identification and basic analysis. However, amplicon-based sequencing does have drawbacks. First, Sanger sequencing is not conducive to the deep coverage needed to detect low-frequency variants. Second, designing PCR primers requires prior knowledge of the viral sequence, which introduces bias and precludes metagenomic analysis. Third, it can be difficult to design primers that produce full-length genomes for all samples, given the high sequence diversity of many viruses. Lastly, degraded samples prevent full-length amplicon production necessary to obtaining whole-genome sequences.

High-throughput sequencing platforms resolve many of the issues of amplicon-based approaches. These platforms, which produce short reads, are better able to capture fragmented or partially degraded samples. They also allow for the ultra-deep sequencing needed to detect low-frequency within-host variants [24]. Sequence independence is essential for studying outbreaks caused by new or unknown pathogens, and for metagenomic analysis. Instead of virus-specific primers, these methods rely on random priming followed by high-throughput sequencing [15, 17, 25]. Combining sequence-independent primer amplification with selective RNase H–based digestion of contaminating RNA (mainly host ribosomal RNA) enables rapid, unbiased deep sequencing of viral samples, as was done during the EBOV epidemic [26].

Other high-throughput sequencing approaches can contribute to viral genomic analysis. Hybrid selection has been used to enrich the viral content of sequencing libraries with high host con-

tamination even after RNase H digestion [18], and is an active area of development [27, 28]. Refining this technology will improve viral genomic analysis during outbreaks, when sample quality may be variable. Other potentially useful technologies still in development include long-read sequencing, which could allow for phasing of variants, and technologies optimized for rapid on-site sequencing. These cheap and portable approaches [29] are useful for rapid diagnostics, but have high error rates that may preclude some detailed genomic analysis.

1.2.2 Recent Advances in Sequencing Methods

Despite the drawbacks of amplicon-based approaches, their importance for sequencing low-titer viruses was highlighted in the 2016 Zika virus outbreak in the Americas [30–33]. Instead of using long amplicons, the amplicon-based sequencing method used in these papers employs a larger number of shorter amplicons (specifically, 35 primer pairs, each creating a ~400-nucleotide amplicon) tiled across the Zika virus genome [34]. Shorter amplicons allow for pathogen identification even in degraded or extremely low titer samples, because large intact regions of the genome need not be present. However, many of the other drawbacks of amplicon-based sequencing remain, such as the inability to reliability detect within-host variants ([31], see Chapter 4).

Hybrid selection, another approach to sequencing samples with low viral content, has also seen recent improvement and more widespread application [18, 35]. This approach relies on small RNA probes that hybridize to the viral genome; washing away any unbound material enriches the proportion of the virus(es) of interest in the sample, thus improving sequencing quality and decreasing cost. Recently, methods have been developed that target a wide variety of viruses at once,

allowing a user to enrich viral content without knowing the specific identity of a disease-causing pathogen [27, 28, 36]. More recently, however, Siddle et al. [37] have developed an algorithm for probe design that maximizes strain diversity while minimizing the number of required probes. Easy probe design is key to more widespread adoption of hybrid selection, and we have already applied the method and resulting probes to both Zika and mumps viruses (see Chapters 4 and 5). New methods to design probe sets reflect both the need to capture newly-identified viruses [38, 39] and the more varied use of target viral capture methods in outbreak and non-outbreak settings.

1.2.3 Sequence Assembly and Alignment

After sequencing, care should be given to the assembly and alignment of genomes. Some of the necessary steps and best practices for processing high-throughput sequencing reads are shown in Figure 1.1 (see also Appendix A). After completing these steps, reads that do not map to the database of possible viruses (Figure 1.1, step 2) can be investigated using one of several taxonomic analysis tools [40–42]. Such reads can be de novo assembled and further investigated using a nucleotide or protein homology search. For a detailed example of how these methods were used to discover a novel flavivirus and two novel rhabdoviruses, see [38, 39]. Alternatively, comprehensive metagenomics pipelines [43, 44] can be used if rapid pathogen identification is the primary goal.

Recombination can affect downstream phylogenetic analysis, so the final sequence alignment should be screened for recombination. Many methods that check for recombination have been compiled into a single software package, RDP4 [46]. As described below, there are alternative phylogenetic tools that should be used when recombination is present, but analysis of recombinant

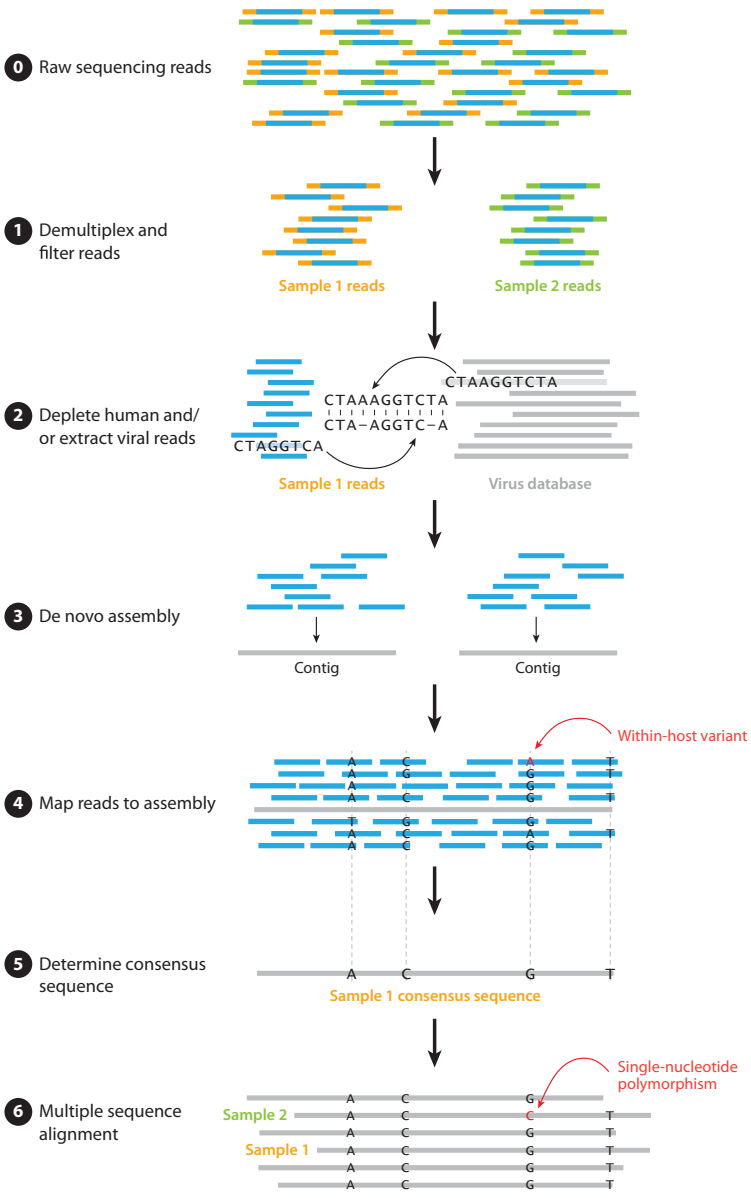


Figure 1.1: Assembly and alignment pipeline for viral reads in heterogeneous samples. High-throughput sequencing reads are (**1**) demultiplexed and filtered for high-quality reads, and (**2**) depleted of host reads [45] and mapped to a database of possible viruses. (**3**) Reads from each sample are de novo assembled, and (**4**) all reads from each sample are mapped onto their own assembly. (**5**) The consensus sequence is determined for each sample and then (**6**) aligned to all other samples using multiple sequence alignment. See Appendix A for available software for each step.

viral sequences is still an area of active development.

1.2.4 Variant Calling

Sequence differences between viral genomes mark the evolutionary history and relationships between samples. Single-base substitutions (single-nucleotide polymorphisms, or SNPs) are the simplest variants. Given high-quality consensus sequences aligned to a reference, it is relatively easy to manually identify SNPs. However, more complex approaches — such as those implemented in packages such as GATK [47] or Samtools [48] — are helpful when samples contain insertions or deletions or when regions of the genome have poor quality, low coverage, or high diversity. Individual SNPs should be annotated — classified as nonsense, missense, or intergenic — and located relative to genes and other genomic elements. Many annotation tools are available online, each requiring only a list of SNPs and an annotated reference genome [49–51].

At this stage, it is also useful to identify variants within individual samples (intrahost single-nucleotide variants, or iSNVs), indicating the presence of multiple viral quasispecies. Powerful tools exist for calling low-frequency variants in heterogeneous viral populations [52, 53]. To avoid calling sequencing errors as iSNVs, we suggest discarding variant calls with fewer than five forward or reverse reads and those for which the number of reads differs greatly between the forward and reverse strands (see Supplemental Methods in [26]). Because PCR errors during library construction can introduce false variants, replicate libraries should be prepared and sequenced whenever possible to confirm the presence of within-host variants at comparable frequencies. The importance of properly filtering within-host variants is extensively discussed in Chapter 4.

1.3 Determining the Origins of an Outbreak

Understanding how and when an outbreak began is critical to curtailing it and to preventing future outbreaks. If an outbreak can be traced to a particular transmission route, steps can be taken to eliminate that route. For example, phylogenetic analysis of human influenza A H5N1 in the 1997 Hong Kong outbreak showed that the virus likely arose through reassortment between an H5N1 virus in terrestrial poultry and a similar virus in quail. This finding led to legislation prohibiting the sale of live quail together with other poultry in Hong Kong [54] and is one of many examples of phylogenetic analysis illuminating the origins of an avian influenza outbreak [55].

Phylogenetic methods all start with the creation of a phylogenetic tree — a reconstruction of the relationship of viral samples to one another — based on nucleotide substitutions in samples from the current outbreak. These phylogenetic relationships can then be used to determine the evolutionary order of sequences and to identify the first cases of an outbreak.

1.3.1 Constructing a Phylogenetic Tree

Phylogenetic trees can be constructed using maximum likelihood [56, 57] or Bayesian [58] approaches. All methods require only a sequence alignment and a nucleotide substitution model. The nucleotide substitution model describes the rate at which one nucleotide is replaced by another and is used to estimate the evolutionary distance between sequences. The model is used in calculating the likelihoods of various possible phylogenetic trees, and it therefore may greatly affect results [59]. A general time-reversible model (typically referred to as a GTR model) is often

used for phylogenetic analyses because it is the most general and makes no assumptions about nucleotide substitution rates or base frequencies [60]. Alternatively, several groups have written statistical software to compare substitution models for a given data set [61].

When constructing or reading trees, it is important to keep confidence values in mind. Confidence in maximum likelihood trees is commonly represented by bootstrap values [62]. Bootstrapping estimates uncertainty by sampling from a dataset with replacement. In this case, the bootstrap value for a node is the proportion of bootstrap trees in which that particular branch topology occurs. Although there is some debate about the accuracy of bootstrap values [63], reporting these values, at least for important nodes, is common practice. Confidence values are built into Bayesian phylogenies and are the posterior probabilities. A Bayesian approach can be thought of as a faster version of a bootstrapped maximum likelihood approach, though the concordance between the two types of confidence values is variable [64].

Both maximum likelihood and Bayesian methods were used to determine the phylogeny of Ebola viruses sequenced during the outbreak [22, 26, 65]. These two methods are often used together to check for agreement: major differences in the resulting trees may suggest a complex evolutionary relationship not fully captured by one or more methods.

1.3.2 Rooting a Phylogenetic Tree

Without further information, a maximum likelihood tree will be unrooted: It will show the relationship of branches relative to one another and the overall topology, but it will not identify the base of the tree or the direction of evolution. It thus cannot be used to identify which samples are

ancestors and which are descendants. Because ancestry is very important to determining the origin of an outbreak, it must be determined by rooting the phylogenetic tree.

There are two primary methods of rooting phylogenetic trees: midpoint rooting and outgroup rooting. Midpoint rooting is done by finding the longest tip-to-tip distance in the tree and setting the root halfway between these tips. This method assumes that evolutionary rates are constant throughout the tree, meaning the root should be equidistant from all tips (it also assumes contemporaneous sampling). As discussed in the next section, this assumption (known as the molecular clock assumption) is often incorrect. Therefore, viral outbreaks are typically rooted by selecting an outgroup — that is, a set of sequences known to be more distantly related than anything else in the tree. This can be comprised of published sequences from previous outbreaks of the same virus, virus sampled from another host species, or a closely related viral species. Outgroup genomes must be distinct from outbreak genomes, but rooting trees using highly divergent sequences can also be problematic [65]. If only outbreak sequences are available, it is also possible to use a particularly divergent cluster of outbreak sequences as the outgroup, if one exists. Once an outgroup is selected, the phylogenetic tree is reconstructed using these additional sequences; the root is the point of divergence between the outgroup sequences and the rest of the tree (Figure 1.2). In some cases the root of the tree is ambiguous, as in the recent EBOV outbreak. Dudas & Rambaut [65] explained how viral substitution rates and linear regression can be used to select the most likely root for a viral outbreak.

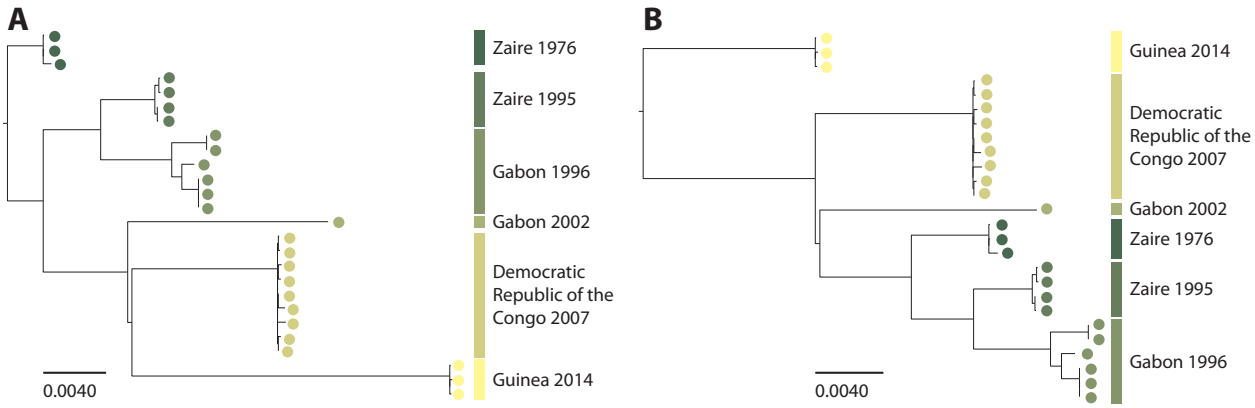


Figure 1.2: Rooting phylogenetic trees. EBOV sequences illustrate the importance of correctly rooting trees. Each point represents one sequence from the outbreak indicated by its color (scale bar = nucleotide substitutions per site). (**A**) Maximum likelihood tree rooted on the Zaire 1976 branch (shown to be the more likely root in [26, 65]). (**B**) The same tree rooted on the Guinea 2014 branch. Interpretation of the ancestral relationships of a single set of samples changes dramatically with root selection.

1.3.3 Estimating the Start Date of an Outbreak

Rooted trees suggest the infection history of an outbreak; outbreak sequences close to the root of the tree (separated by fewer nodes) came earlier in the outbreak. If sampling dates are available, branch lengths can be converted from units of nucleotide substitutions to units of time, which can be used to date the true origin of the outbreak. This is done using a strict molecular clock model, which assumes that nucleotide substitutions accumulate at a constant rate [66]. The number of substitutions on each branch of a tree with dated tips can be used to estimate the nucleotide substitution rate, which then can be used to extrapolate backward to the date of origin of a particular outbreak strain [67]. Maximum likelihood methods [68] can calculate substitution rates given a phylogenetic tree and sampling dates.

Although it is a helpful simplification, the strict molecular clock does not always accurately

model real viral evolution; evolutionary rates can vary over time, over space, or between different branches. To address this, more flexible models have been developed that allow for variation in the substitution rate over time [69]. The Bayesian Evolutionary Analysis by Sampling Trees (BEAST) package [70] implements a Bayesian Markov chain Monte Carlo method to determine changing substitution rates over time; this is referred to as a relaxed molecular clock. This framework can be used to coestimate the phylogeny and divergence times given sequence data and sampling dates. During the EBOV outbreak, BEAST was used to estimate when outbreak viruses split from lineages documented in other outbreaks [65] and to estimate the date of entry of the virus into Sierra Leone from Guinea [26, 71].

1.3.4 Determining the Cause of an Outbreak

The same phylogenetic methods can be used to determine the type(s) of transmission causing an outbreak (human-human or animal-human), but this analysis requires sequences from appropriate hosts and/or time periods (Figure 1.3). For example, analysis of EBOV patient samples showed that there was substantial genetic variation between EBOV outbreaks, but limited variation within each outbreak. This suggested that the virus evolves separately in an animal reservoir, and that a single zoonotic transmission was responsible for the start of each outbreak. This hypothesis was supported by the divergence times calculated by BEAST: The lineages from the two most recent outbreaks diverged from a common ancestor significantly before the start of either outbreak [26].

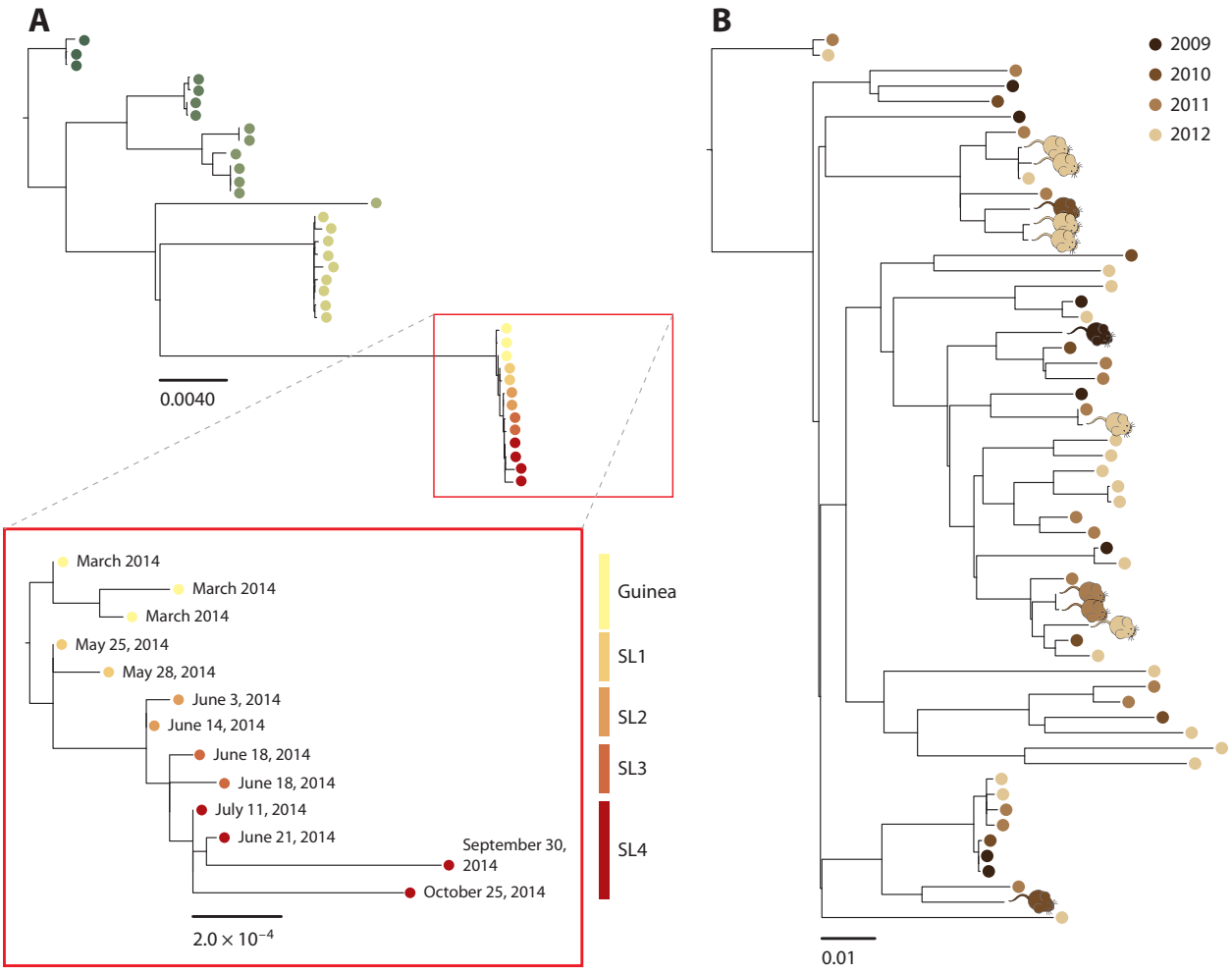


Figure 1.3: Tree topology illuminates the nature of an outbreak. (**A**) EBOV tree. Sequences from previous outbreaks are colored in distinct shades of green (see Figure 1.2). Selected sequences from the Sierra Leone outbreak highlight the low diversity within the outbreak, and the development of new clades (SL1–4, defined in [19, 26]) from a single recent ancestor. This topology suggests that each outbreak began with a single zoonotic transmission but was subsequently sustained by human-to-human transmission. (**B**) Lassa virus (LASV) tree containing S segment sequences from both human (circle nodes) and *Mastomys natalensis* (rodent nodes) hosts. Samples are from Sierra Leone [72], where LASV is endemic. Sequences do not cluster by time or by host, indicating frequent animal-to-human transmission and a lack of discrete outbreaks.

1.3.5 Challenges in Estimating the Origin of an Outbreak

Although phylogenetic tools have been used successfully to understand many viral outbreaks, significant challenges remain in correctly establishing the origin of an outbreak. A detailed review of current challenges in phylogenetic methods can be found in [73]. Here we highlight those challenges particularly relevant for determining the origin of a viral outbreak.

First, although relaxed molecular clocks allow for some rate variation, current models may still fail to capture the full variation in evolutionary rates. For example, an analysis for pandemic HIV-1 group M found that the time to the most recent common ancestor varies significantly when subtypes are analyzed separately compared with jointly, perhaps because closely related viral lineages have different substitution rates [74].

Additionally, the methods described above cannot account for recombination that occurs in many viruses, because the ancestry of these viruses cannot be represented by a simple branching process. Instead, different parts of a single sample's genome can be the product of different genealogical trees and are better modeled by a phylogenetic network or ancestral recombination graph that allows for complicated evolutionary relationships [75–77]. Because many phylogenetic tools cannot account for recombination, it is important to restrict analysis to parts of the viral genome or tree where recombination is limited. Important advances in the field will come from continued development of phylogenetic tools that can incorporate viral recombination.

Lack of data also poses significant barriers to analyzing many viral outbreaks. For example, lack of sampling dates essentially rules out divergence time estimates, and lack of informative out-

group sequences — perhaps due to limited past sequence data, or because a zoonotic reservoir has yet to be determined — prevents accurate rooting of a phylogenetic tree. Nonrandom sampling over time or space may significantly bias results [73]. Finally, understanding the ecological factors leading to an outbreak at a particular place and time requires detailed surveys of the outbreak location, both during and before its start [78, 79]. Without detailed epidemiological surveys, it may be impossible to determine the index case of a viral outbreak.

1.4 Viral Disease Transmission

Understanding the spread of the virus, including the mechanism, speed, and direction, is essential to controlling a viral outbreak. In many cases, this is done with epidemiological modeling. During the EBOV epidemic, many groups used case counts to estimate epidemiological parameters and the eventual size of the epidemic [80–90]. The varied approaches taken by these groups illustrate that there is no standard way to parameterize and use these epidemiological models. Additionally, in the absence of very detailed contact tracing and other epidemiological metrics, these models often cannot capture the complexity of an outbreak.

Even without whole-genome sequencing, studying different viral strains as they move through time and space can be used to determine transmission patterns, especially for viruses with distinct subtypes, like HIV or influenza virus. However, this type of analysis does not always have adequate resolution to answer important questions about transmission routes. In the case of possible HIV transmission from a dental procedure, as described in Section 1.1, both contact tracing and molec-

ular methods failed to prove a link between dentist and patient: contact tracing led to the dentist, but it was not conclusive because there was no evidence of shared bodily fluids. Similarly, two individuals with the same HIV subtype do not indicate a direct transmission link. Sequencing of the viruses from the patient, dentist, and several local individuals finally provided significant evidence for direct transmission [14].

Using genetic data to reconstruct transmission routes is also especially important for post-outbreak cases — those who are infected after all known transmission routes have died out. For example, an individual contracted EBOV 68 days after human-to-human transmission linked to the 2014–2016 EBOV epidemic was declared to have ended [91]. Because the infected individual had no known Ebola virus-positive contacts, and because molecular diagnostic methods could not differentiate between outbreak lineages, whole-genome sequencing was needed to identify the most likely source of infection.

As shown by these examples, viral genomic sequencing can be a vital tool for detailed transmission reconstruction when contact tracing data is missing or uninformative. Genetic data can also be used to identify general transmission patterns and to estimate epidemiological parameters.

1.4.1 Transmission Reconstruction

Reconstructed viral transmission routes can reveal modes of transmission that are important for containment and prevention [7]. Depending on the depth of sampling, rooted phylogenetic trees either coarsely or in detail correspond to the transmission history of the virus [92–95] (Figure 1.4).

Formal methods have been developed to reconstruct transmission chains from genetic data

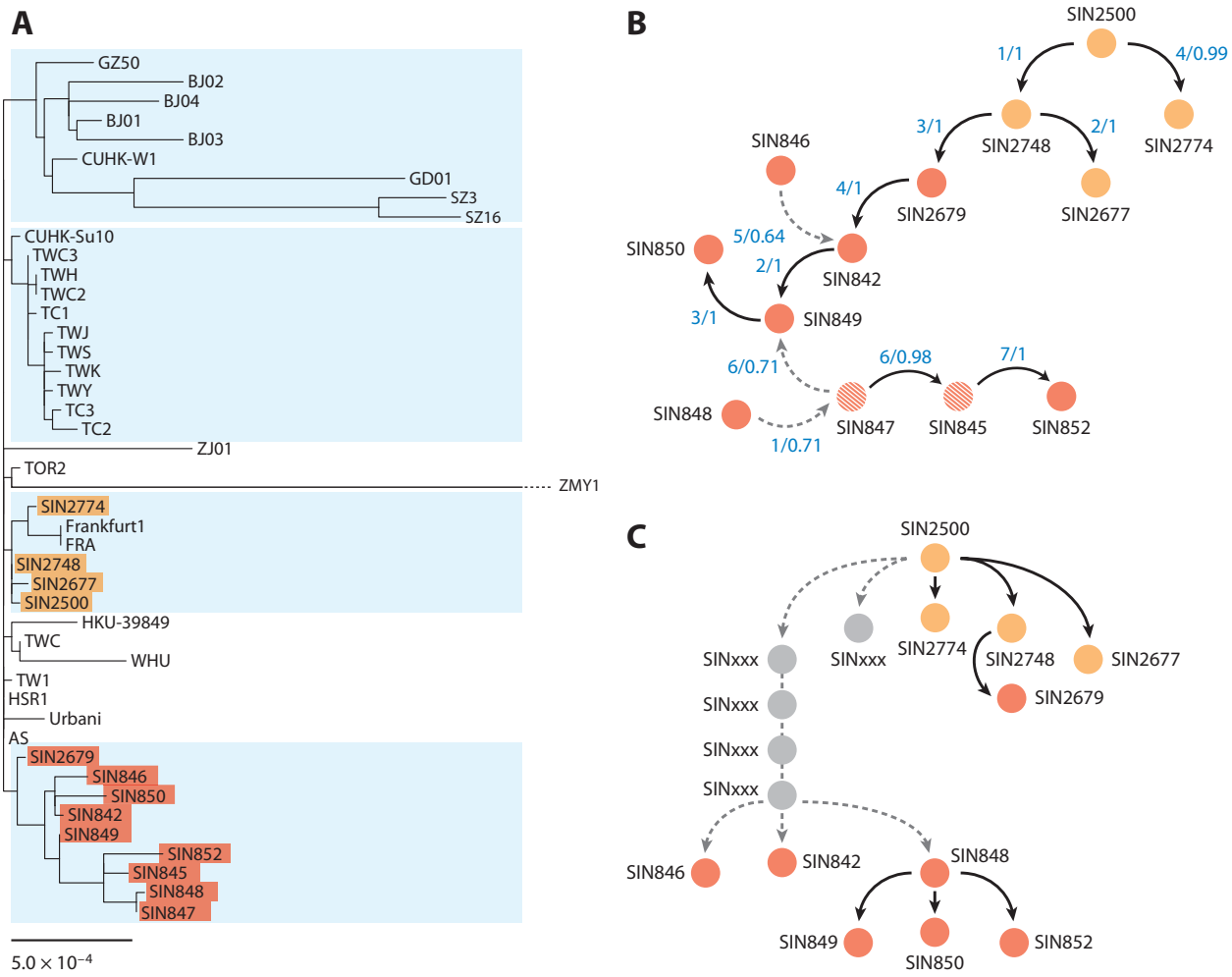


Figure 1.4: Determining virus transmission. Transmission during the 2003 SARS outbreak in Singapore, for which both sequence and contact tracing data are available. (**A**) Maximum likelihood tree rooted on TOR2, the earliest reported case. The four major branches (blue boxes) roughly correspond to geographic origin (top to bottom: China, Taiwan, Singapore, Singapore). (**B**) Transmission tree reconstruction from sequences and sample collection dates (created using outbreaker [96]); generation time: gamma distribution with mean = 8.4 and sd = 3.8, based on values from [2]). Red and yellow circles correspond to the two Singapore clades identified in (A); lined red circles are samples with only sequence data (no contact tracing). Arrows are labeled with (number of SNPs between samples) / (posterior probability of transmission). (**C**) Transmission tree created during the SARS outbreak by contact tracing, as reported by [97]. Gray circles are unreported cases assumed to be part of the transmission chain. Comparison of panels (A–C) shows that the three methods generate similar relationships between samples.

[96]; several of these methods combine genetic and epidemiological data (e.g., sampling dates and locations) into a single likelihood function that is used to sample possible transmission trees [92, 96, 98, 99]. Jombart et al. [96, 100] have developed an R package that constructs transmission trees from genetic and any available epidemiological data (Figure 1.4B).

It has recently been recognized that within-host genomic data constitute an essential component of phylogenetic analysis and transmission tree reconstruction [67, 99, 101]. One challenge of incorporating this type of data is that it requires an understanding of the characteristic within-host dynamics for a given virus before it can be used to effectively inform statistical and epidemiological models. Specifically, it is important to know the underlying viral mutation rate and the typical within-host nucleotide substitution rate, as well as how much diversity is transmitted during an infection event (the bottleneck size). Because studying within-host dynamics requires both high sequencing depth and longitudinal sampling, limited information exists for most viruses. Within-host studies are most common in well-studied chronic viral infections such as HIV infection [102, 103], although similar studies in other viruses are beginning to appear [104]. In the same vein, the average size of the transmission bottleneck is known in HIV [105] but is still under investigation in most other viruses. Deep sequences of Ebola viruses published during the 2014–2016 outbreak suggest that the bottleneck size is greater than one [19, 106], but more precise estimates are still needed. Within-host viral dynamics studies, along with the development of robust phylogenetic methods that incorporate within-host variation, are a crucial next step in outbreak research.

1.4.2 Recent Advances in Transmission Reconstruction

The last few years have, in fact, produced a number of new methods that incorporate within-host variation into transmission analysis [107–109]. Allowing for within-host variation in transmission models was an essential step in fully connecting transmission and phylogenetic trees; previously, transmission trees were often estimated without phylogenetic information (such as in [96]) or transmission was reconstructed given a fixed phylogenetic relationship between samples [99, 110]. Consequently, many recent studies [107–109] feature simultaneous sampling of phylogenetic and transmission trees. Simultaneous sampling over both types of trees allows these methods to better account for missing samples, thereby improving transmission reconstruction. Despite these improvements, dealing with unsampled cases remains a challenge in the field, and several recent papers have addressed and suggested potential solutions to the problem [99, 111, 112].

Another active area of development is relaxing the assumption — made in all three of the simultaneous sampling methods mentioned above — that the transmission bottleneck size is equal to one. Within-host data may play an important role in this, as previously suggested by Worby et al. [113]. Even with extensive within-host data, it is important to remember that transmission reconstruction is ultimately limited by the mutation rate of the virus [114], since it is difficult to infer patterns from identical sequences. This suggests the important role of epidemiological data in transmission inference: as discussed at length in Chapter 5, an approach combining genomic and epidemiological data, or genomic epidemiology, has the potential to be more powerful than one using either datatype alone.

1.4.3 Calculating the Reproduction Number

The basic reproduction number (R_0) — the number of secondary cases from a single infection — is a useful measure of the infectivity of a pathogen and is usually estimated from epidemiological models. However, this number can also be estimated from a detailed transmission chain or from genomic data [115, 116]. This value often frames the discussion about containment for a disease outbreak and can be used to predict outbreak dynamics and eventual size in the presence or absence of various control measures (for its use in the EBOV outbreak, see [82, 117]). Calculation of epidemiological parameters such as R_0 is part of the new and growing field of phylodynamics, the study of infectious disease behavior that arises from a combination of evolutionary and epidemiological processes [102]. Incorporating epidemiological metadata can enhance genetic analysis, and vice versa, in outbreak situations.

1.4.4 Challenges in Understanding Transmission

Although joint evolutionary and epidemiological analysis has greatly advanced the field of outbreak investigation, there are still challenges associated with determining the route and rate of viral spread. Major hurdles include sampling bias and the difficulty of allowing for spatial and temporal complexity in phylodynamic models [73]. For example, Lloyd-Smith et al. [118] highlighted problems with using the same reproduction number for all individuals and the resulting implications for outbreak control.

1.5 Viral Evolution

Beyond epidemiological questions, sequencing is well suited to studying viral evolution during an outbreak. Of particular interest are the appearance and spread of mutations that affect viral fitness or virulence, or those that allow the virus to evade vaccines or immune responses. That said, it is important to remember that it is often impossible to draw meaningful conclusions from outbreak data without experimental validation. With this in mind, we discuss viral mutation and substitution rates and review metrics of selection in viral populations. We focus on metrics relevant to understanding an outbreak. A more detailed discussion of viral mutation and substitution rates in particular can be found in the review by Duffy et al. [119].

1.5.1 Mutation and Substitution Rates

The mutation rate is a major determinant of the overall rate of evolutionary change during and between outbreaks; it is the number of genetic mutations that occur per viral genome replication. It is largely determined by a virus's biological properties, such as the fidelity of its polymerase, the speed at which it replicates its own genome, and whether the genome is RNA or DNA [119]. In general, RNA viruses mutate fastest and DNA viruses slowest. The mutation rate must be measured experimentally because natural selection affects the number of mutations identified in genetic data. For mutation rate estimates for a number of specific viruses, see the reports by Drake [120] and Drake & Hwang [121].

The mutation rate should not be confused with the nucleotide substitution rate, which is the

rate at which nucleotide substitutions accumulate in a viral lineage. This rate is determined by the mutation rate and by other factors, including natural selection and the effective viral population size. This is the rate most commonly discussed in an outbreak situation, both because it can be calculated from sequence data and because it can be used to understand selective pressures on a viral population during an outbreak. For example, it may be useful to compare the substitution rate within an outbreak to that in a zoonotic reservoir. The virus should have the same intrinsic mutation rate in both hosts, so differences in substitution rate could be due to selection.

Whereas calculating the viral mutation rate requires careful experimentation, the substitution rate can be calculated given a phylogenetic tree and sampling dates. This can be done with maximum likelihood methods [68] or Bayesian methods [70]. Bayesian methods such as BEAST are more statistically rigorous than most maximum likelihood implementations because they can allow the substitution rate to vary between branches, but they are computationally intensive [119]. For approximate substitution rates for various viruses, see the compilation by Jenkins et al. [122].

The caveat to all of these methods is that they assume all substitutions are fixed in the population. Because many mutations on recent branches are mildly deleterious and will disappear from the population over time, the substitution rate for any tree containing recent samples may be artificially high. This is common during outbreaks, when a majority of viral genomes may be terminal branches.

1.5.2 Rate-Based Tests for Selection

During an outbreak, identification of variants that confer a fitness benefit to the virus is a key concern. Variants that affect fitness are by definition under selection in a population. A well-known statistical test for selection is the d_N/d_S (also called K_a/K_s or ω) test (Figure 1.5). This test compares the rates of synonymous and nonsynonymous substitutions in some region of the genome. In the absence of selection, these two rates should be the same (corrected for the frequency of each type of site). Nonsynonymous substitutions are more likely to affect the resulting protein and are therefore more likely to influence fitness. Therefore, the ratio of synonymous to nonsynonymous changes can indicate selection: $d_N/d_S > 1$ suggests positive selection, and $d_N/d_S < 1$ suggests negative or purifying selection.

This test requires only a codon alignment and is often applied to viral sequences to identify domains or genes under selection [123, 125, 126]. However, it was originally developed to analyze sequences from divergent species, not to detect selection within a single population [127]. The d_N/d_S ratio is not applicable within single populations, and the results obtained from this test are often misleading when applied to microbes [128]. Therefore, although this ratio can still be used to analyze sufficiently divergent, separately evolving outbreaks, users should be wary of using this test to detect selection within a single outbreak population.

If data from other outbreaks are available, the d_N/d_S statistic can be used to identify sites or regions under selection. During the EBOV epidemic, several groups found $d_N/d_S > 1$ in the gene encoding the glycoprotein (GP) — or, more specifically, the disordered, mucin-like region of GP —

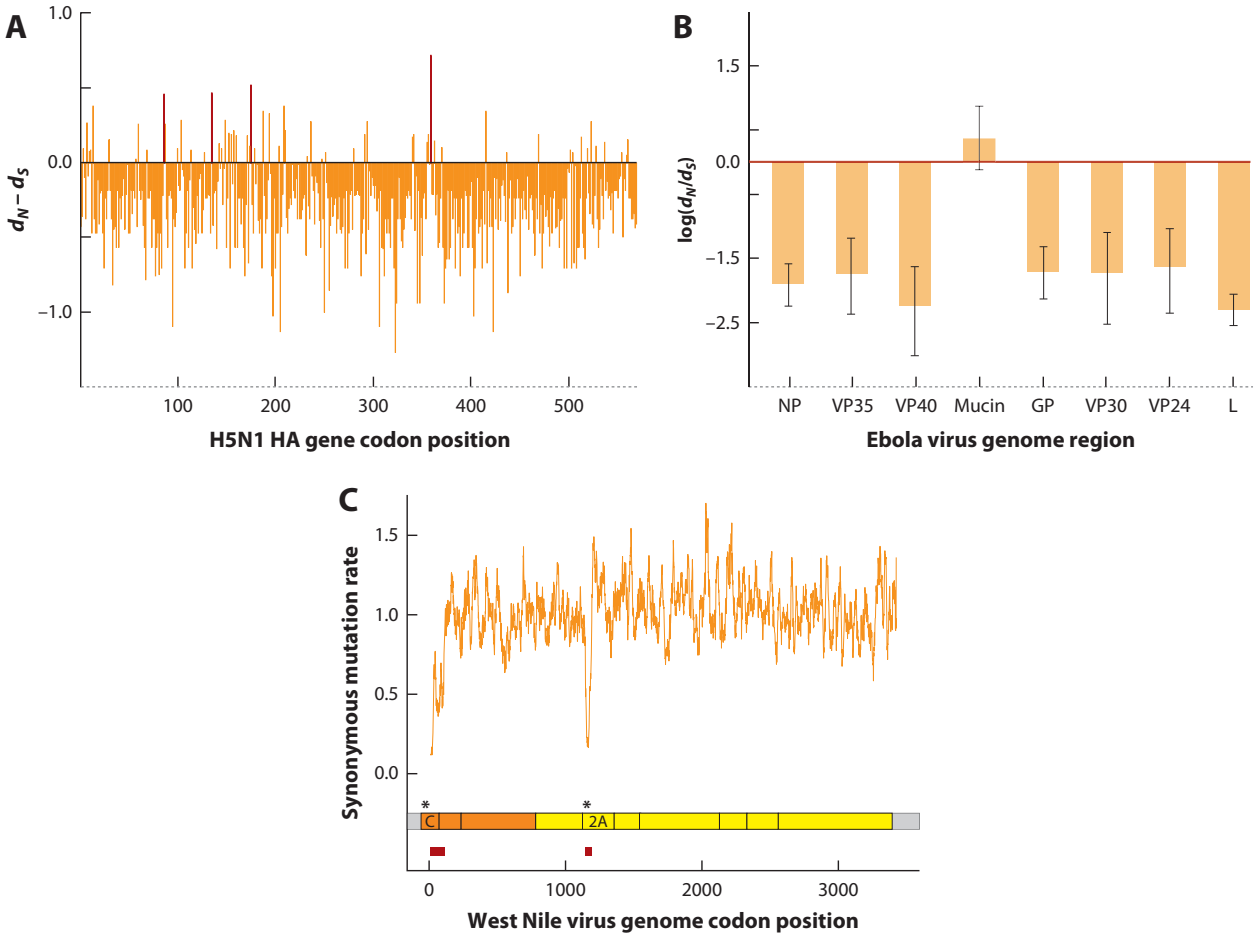


Figure 1.5: Rate-based tests for selection in viruses. Various tests identify signals of selection at different genomic scales. (**A**) $d_N - d_S$ scores for every codon in the hemagglutinin (HA) gene of H5N1 influenza A (calculated using [123]). The highest $d_N - d_S$ scores (red) indicate codons most likely under positive selection. (**B**) $\log(d_N/d_S)$ for each EBOV gene and for the mucin-like region of the glycoprotein (GP) (values from [19]). (**C**) Synonymous constraint for every codon position in the West Nile virus genome (sliding window = 20 nucleotides) [124]. Red bars mark regions of excess constraint. Asterisks mark two known RNA structural elements (orange = structural proteins, yellow = non-structural proteins), a hairpin in the capsid gene and a pseudoknot element within non-structural protein 2A. Panel (C) adapted with permission from [124]

when including sequences from all known EBOV outbreaks [19, 129] (Figure 1.5B). This is unsurprising because GP is the envelope protein: as the only surface-exposed protein on the viral particle, GP is the target of host cell antibodies. Because of this biology, it is suspected that GP undergoes diversifying selection or relaxed purifying selection as a response to host immune pressure [130].

In general, comparing nonsynonymous to synonymous mutations between sites or between timescales (i.e., comparing inter- and intrahost substitutions, as in [72]) can be used to suggest regions under selection. It is possible to identify selection using only synonymous substitutions by looking for regions of excess synonymous constraint in a virus [124] (Figure 1.5C). In viruses, many protein-coding regions contain overlapping or embedded functional elements. Because any type of substitution may disrupt an overlapping element, these regions are often characterized by an unusually low synonymous substitution rate. This method was used during the EBOV epidemic to find a constrained region in a known editing site.

1.5.3 Other Tests for Selection

Tests based on mutation frequency spectra and tree topology can also be used to identify selection in single populations. The basic principle behind these methods is that selective pressure on a population leaves a distinct mark on overall genetic diversity and tree symmetry [102, 131, 132]. Statistics that test for selection, or non-neutrality, using these principles include Tajima's D, Fu and Li's D, and tree-imbalance metrics [133].

Tajima's D is a statistic that compares the average number of pairwise differences between sequences to the total number of variable sites within the set of sequences [134]. A negative value of

Tajima's D indicates an excess of low-frequency polymorphisms compared to a neutral model, and may be due to purifying selection or population size expansion. Conversely, a population size reduction (or bottleneck) and balancing selection keep variants at intermediate frequencies, which translates to a positive value for Tajima's D. Fu and Li's D applies this idea to the phylogeny of these sequences and compares the number of mutations on newer, external branches to the total number of mutations on all branches [135]. As with Tajima's D, a basic understanding of population genetics can be used to interpret the result of this test. For example, under purifying selection, there is likely to be an excess of mutations on external branches because deleterious mutations are generally purged from the population before they can be passed on. Three measures of tree imbalance are also commonly used to test for selection: B_I [136], the cherry count [137], and Colless's tree imbalance index [138]. More asymmetry than expected in a phylogenetic tree suggests non-neutral evolution. These methods have been successfully used to understand selection in HIV, influenza virus, and other viruses [133, 139].

The major drawback of both frequency spectra and tree imbalance methods is that it can be hard to differentiate between selection and epidemiological effects such as changing population size. For example, a very negative Tajima's D or Fu and Li's D can be due to exponential growth rather than non-neutral evolution. Drummond & Suchard [133] addressed this problem by incorporating a demographic model when analyzing three RNA virus data sets and showed that it is possible to use these tests to identify selective pressure on viral populations.

1.5.4 Challenges in Viral Evolution

Detecting selection in viruses is challenging because most statistical methods have been created for the comparison of divergent populations or species, rather than for analysis of a single population that may be rapidly evolving and expanding. Additionally, viruses are very biologically diverse and have highly variable mutation and substitution rates. This makes it difficult to use the same selection tests for all viruses. For example, slow-mutating viruses, which usually have low substitution rates, may require extended sampling to achieve the population diversity needed to identify evolutionary trends. Unfortunately, not all viruses and data sets make good subjects for evolutionary analysis, and even when they do, the results may be relatively uninformative or uninteresting.

1.6 Functional Variation

One important role of genomic data is to inform experimental studies, which are necessary for understanding the biology of pathogenic viruses. In many cases, genetic analysis of outbreak sequences generates hypotheses about particular regions of a virus that may play a role in transmission or pathogenesis. Validating or refuting these hypotheses experimentally leads to a more complete picture of the virus, which may directly inform treatment and prevention measures or be used to improve epidemiological and evolutionary models.

Immediately after sequencing, viral samples can be used to answer one pressing question: whether one or more mutations have impaired the ability of clinical diagnostic tools to detect the virus. This is particularly a concern for the real-time PCR assays commonly used for viral detection,

because SNPs and other mismatches in primer binding sites have been shown to greatly reduce assay performance [140]. For example, during the EBOV epidemic, various groups periodically compared the most up-to-date list of mutations in the EBOV genome with recognition sites for diagnostic probes, as well as for existing and candidate therapeutics [141–143]. Mutations in the binding regions of diagnostics or therapies should be carefully tested experimentally to ensure that binding still occurs.

Broadly, the results of phylogenetic and evolutionary analyses can be used to identify variants most likely to have a functional effect on the virus. For example, clade-defining mutations — mutations shared by large clusters on a phylogenetic tree — are prime candidates for experimentation. These mutations may have fixed within a cluster of samples simply by genetic drift and patterns of transmission, but they could also represent sites under strong positive selection. For example, genomic analysis of EBOV sequences demonstrated the presence of four viral lineages circulating in Sierra Leone, each defined by one to four deviations from the reference genome, that rose to prevalence in the population at some point during the outbreak [19, 26, 144, 145]. Because of their prominence, these mutations were targeted for experimental study soon after the outbreak started [146].

Variants or genomic regions identified by the evolutionary analyses described in the previous section should also be considered for experimental testing. Analyses suggesting that the glycoprotein in EBOV might be under selection are an excellent example of how genomic analyses were not able to definitively classify selective pressures, but were able to identify the most promising region for functional validation.

Another common question during an outbreak is whether mutations correlate with clinical outcomes. Therefore, before conducting experiments, it may be informative to explore mutations in relation to clinical and other types of data. This requires additional data, such as information about symptoms, survival, or viral load. Correlations cannot prove causation but can be used to refine a set of mutations for experimental analysis and to suggest a function or mechanism that can be tested experimentally.

1.7 Remaining Challenges in Viral Genomics

Genomic analysis can answer many urgent questions during an ongoing viral outbreak, including ones related to where the outbreak originated, how the virus is transmitted, and how the virus might be evolving. This was successfully done during the EBOV outbreak, and the same techniques are being applied to past and ongoing outbreaks of other infectious diseases. However, all of these analyses are limited by the quality and availability of data.

Data quality may be improved by updated sample preparation and sequencing methods. These methods are especially important for viruses that are present at low titer, such as Zika virus. To best utilize sequencing data, many of the analyses discussed could be further refined with better information about the virus itself [147]. For example, biological investigation of the evolutionary and transmission processes unique to specific viruses would improve the quality of many within-outbreak analyses.

Although these technical challenges remain, logistical issues are a major barrier to effective

outbreak response because phylogenetic methods depend on specific types of data: determining the time an outbreak started is difficult if a suitable outgroup is not available, reconstructing transmission is impeded by inaccurate or missing sample dates, and all of these techniques are limited by sparse sampling during an outbreak. Missing data have been a major challenge in viral genomics largely because the usefulness of real-time sample collection and sequencing for outbreak control was not recognized until recently. Now, with easier, cheaper sequencing and the development of computational methods to harness that sequence data, it should be evident that genomic data will be a powerful tool in understanding and controlling future outbreaks.

Even when data are collected, decentralized sample collection and analysis mean that those data may not be readily available. This problem was highlighted in the EBOV epidemic, during which many different groups were conducting studies all over Western Africa, and the data did not always become immediately available. Sharing outbreak data is a necessary component of an efficient response [148]. Another lesson from the EBOV epidemic is the usefulness of extensive collaboration. Many of the techniques discussed in this review are computationally intensive (BEAST, for example), and deep sequencing of isolates is very expensive; both require substantial technical expertise. Large collaborations and shared resources and data seem to be the best ways to respond quickly to an outbreak situation. Although this has not been the normal approach of many research groups, informal collaborations, such as the online forum Virological (http://virological.org) and open-source project Nextstrain (http://www.nextstrain.org), are already influencing how we respond to outbreaks.

Genomic analysis is a powerful tool for understanding, and therefore combating, viral out-

breaks. The field is at a fundamental transition point, supported by recent improvements in viral sequencing and analysis. The biggest immediate hurdle is data collection sufficient to enable full utilization of these methods.

In order to best prepare for future viral outbreaks, we must facilitate collection and rapid sequencing of viral samples. This requires the development of cheap and effective viral diagnostics for use in the field and a continued effort to promote data-sharing and collaboration among viral researchers. Additionally, a clear goal for the future of viral genomics is a more complete integration of epidemiological and genetic data in statistical methods. In taking these steps, we can develop the capacity to deal with viral and microbial outbreaks and other threats related to infectious disease.

1.8 Acknowledgements

We thank T. Bedford, A. Lin, B. MacInnis, C. Matranga, D. Nosamiefan, D. Park, A. Piantadosi, H. Metsky, S. Ye, and N. Yozwiak for their helpful comments.

1.9 References

[1] Wohl, S., Schaffner, S. F., and Sabeti, P. C. Genomic analysis of viral outbreaks. *Annual Review of Virology*, 2016.

[2] Lipsitch, M., Cohen, T., Cooper, B., et al. Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300(5627):1966–1970, 2003.

[3] Kiss, I. Z., Green, D. M., and Kao, R. R. Disease contact tracing in random and clustered networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570):1407–1414, 2005.

[4] Porco, T. C., Holbrook, K. A., Fernyak, S. E., et al. Logistics of community smallpox control through contact tracing and ring vaccination: a stochastic network model. *BMC public health*, 4:34, 2004.

[5] Kretzschmar, M., van den Hof, S., Wallinga, Jacco, and van Wijngaarden, J. Ring vaccination and smallpox control. *Emerging Infectious Diseases*, 10(5):832–841, 2004.

[6] Klinkenberg, D., Fraser, C., and Heesterbeek, H. The effectiveness of contact tracing in emerging epidemics. *PLoS ONE*, 1:e12, 2006.

[7] Grad, Y. H. and Lipsitch, M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biology*, 15(11):538, 2014.

[8] Centers for Disease Control and Prevention. Guidance for clinicians on the use of RT-PCR and other molecular assays for diagnosis of influenza virus infection, 2018. URL http://www.cdc.gov/flu/professionals/diagnosis/molecular-assays.htm.

[9] Harper, S. A., Bradley, J. S., Englund, J. A., et al. Seasonal influenza in adults and children–diagnosis, treatment, chemoprophylaxis, and institutional outbreak management: clinical practice guidelines of the Infectious Diseases Society of America. *Clinical Infectious Diseases*, 48(8):1003–1032, 2009.

[10] Russell, C. A., Jones, T. C., Barr, I. G., et al. The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874):340–346, 2008.

[11] Rota, P. A., Oberste, M. S., Monroe, S. S., et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300(5624):1394–1399, 2003.

[12] Wang, D., Urisman, A., Liu, Y.-T., et al. Viral discovery and sequence recovery using DNA microarrays. *Plos Biology*, 1(2):E2, 2003.

[13] World Health Organization. PCR primers for SARS developed by WHO Network Laboratories, 2003. URL http://www.who.int/csr/sars/primers/en/.

[14] Ou, C. Y., Ciesielski, C. A., Myers, G., et al. Molecular epidemiology of HIV transmission in a dental practice. *Science*, 256(5060):1165–1171, 1992.

[15] Djikeng, A., Halpin, R., Kuzmickas, R., et al. Viral genome sequencing by random priming methods. *BMC genomics*, 9:5, 2008.

[16] Ninomiya, M., Ueno, Y., Funayama, R., et al. Use of illumina deep sequencing technology to differentiate hepatitis C virus variants. *Journal of clinical microbiology*, 50(3):857–866, 2012.

[17] Malboeuf, C. M., Yang, X., Charlebois, P., et al. Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic acids research*, 41(1): e13, 2013.

[18] Matranga, C. B., Andersen, K. G., Winnicki, S., et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biology*, 15(519), 2014.

[19] Park, D. J., Dudas, G., Wohl, S., et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526, 2015.

[20] Flint, M., Goodman, C. H., Bearden, S., et al. Ebola Virus Diagnostics: The US Centers for Disease Control and Prevention Laboratory in Sierra Leone, August 2014 to March 2015. *The Journal of Infectious Diseases*, 212 Suppl 2:S350–8, 2015.

[21] Towner, J. S., Sealy, T. K., Ksiazek, T. G., and Nichol, S. T. High-throughput molecular detection of hemorrhagic fever virus threats with applications for outbreak settings. *The Journal of Infectious Diseases*, 196 Suppl 2:S205–12, 2007.

[22] Baize, S., Pannetier, D., Oestereich, L., et al. Emergence of Zaire Ebola Virus Disease in Guinea — Preliminary Report. *New England Journal of Medicine*, 371(15):1418–1425, 2014.

[23] Quick, J., Loman, N. J., Duraffour, S., et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–233, 2016.

[24] Watson, S. J., Welkers, M. R. A., Depledge, D. P., et al. Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120205, 2013.

[25] Reyes, G. R. and Kim, J. P. Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Molecular and cellular probes*, 5(6):473–481, 1991.

[26] Gire, S. K., Goba, A., Andersen, K. G., et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[27] Briese, T., Kapoor, A., Mishra, N., et al. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *Mbio*, 6(5):e01491–15, 2015.

[28] Wylie, T. N., Wylie, K. M., Herter, B. N., and Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Research*, 25(12):1910–1920, 2015.

[29] Greninger, A. L., Naccache, S. N., Federman, S., et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome medicine*, 7(1):99, 2015.

[30] Quick, J., Grubaugh, N. D., Pullan, S. T., et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, 12(6):1261–1276, 2017.

[31] Metsky, H. C., Matranga, C. B., Wohl, S., et al. Zika virus evolution and spread in the Americas. *Nature*, 546(7658):411–415, 2017.

[32] Grubaugh, N. D., Ladner, J. T., Kraemer, M. U. G., et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature*, 546(7658):401–405, 2017.

[33] Faria, N. R., Quick, J., Claro, I. M., et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*, 546(7658):406–410, 2017.

[34] Worobey, M., Watts, T. D., McKay, R. A., et al. 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature*, 539(7627):98–101, 2016.

[35] Depledge, D. P., Palser, A. L., Watson, S. J., et al. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS ONE*, 6(11):e27805, 2011.

[36] Chalkias, S., Gorham, J. M., Mazaika, E., et al. ViroFind: A novel target-enrichment deep-sequencing platform reveals a complex JC virus population in the brain of PML patients. *PLoS ONE*, 13(1):e0186945, 2018.

[37] Siddle, K. J., Metsky, H. C., Gladden-Young, A., et al. Capturing diverse microbial sequence with comprehensive and scalable probe design. *bioRxiv*, 2018. doi: 10.1101/279570. URL https://www.biorxiv.org/content/early/2018/03/12/279570.

[38] Stremlau, M. H., Andersen, K. G., Folarin, O., et al. Discovery of Novel Rhabdoviruses in the Blood of Healthy Individuals from West Africa. *PLOS Neglected Tropical Diseases*, 9(3): e0003631, 2015.

[39] Kapoor, A., Kumar, A., Simmonds, P., et al. Virome Analysis of Transfusion Recipients Reveals a Novel Human Virus That Shares Genomic Features with Hepaciviruses and Pegiviruses. *Mbio*, 6(5):e01466–15, 2015.

[40] Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, 2011.

[41] Wood, D. E. and Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

[42] Segata, N., Waldron, L., Ballarini, A., et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814, 2012.

[43] Naccache, S. N., Federman, S., Veeraraghavan, N., et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*, 24(7):1180–1192, 2014.

[44] Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature Biotechnology*, 29(5):393–396, 2011.

[45] MacConaill, L. and Meyerson, M. Adding pathogens by genomic subtraction. *Nature Genetics*, 40(4):380–382, 2008.

[46] Martin, D. P., Murrell, B., Golden, M., Khoosa, A., and Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1, 2015.

[47] McKenna, A., Hanna, M., Banks, E., et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[48] Li, H., Handsaker, B., Wysoker, A., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[49] Cingolani, P., Platts, A., Wang, L. L., et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80, 2012.

[50] Chen, Y., Cunningham, F., Rios, D., et al. Ensembl variation resources. *BMC genomics*, 11:293, 2010.

[51] Wang, K., Li, M., and Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164, 2010.

[52] Yang, X., Charlebois, P., Macalalad, A., Henn, M. R., and Zody, M. C. V-Phaser 2: variant inference for viral populations. *BMC genomics*, 14:674, 2013.

[53] Koboldt, D. C., Chen, K., Wylie, T., et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 2009.

[54] Guan, Y., Peiris, J. S. M., Lipatov, A. S., et al. Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. *PNAS*, 99(13):8950–8955, 2002.

[55] Lei, F. and Shi, W. Prospective of Genomics in Revealing Transmission, Reassortment and Evolution of Wildlife-Borne Avian Influenza A (H5N1) Viruses. *Current genomics*, 12(7):466–474, 2011.

[56] Guindon, S. and Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696–704, 2003.

[57] Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[58] Huelsenbeck, J. P. and Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[59] Posada, D. and Crandall, K. A. Selecting the Best-Fit Model of Nucleotide Substitution. *Systematic biology*, 50(4):580–601, 2001.

[60] Tavare, S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17, 1986.

[61] Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8):772, 2012.

[62] Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39:783–791, 1985.

[63] Efron, B., Halloran, E., and Holmes, S. Bootstrap confidence levels for phylogenetic trees. *PNAS*, 93(23):13429–13434, 1996.

[64] Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F., and Douzery, E. J. P. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol*, 20(2):248–254, 2003.

[65] Dudas, G. and Rambaut, A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. *PLoS Currents*, pages 1–9, 2014.

[66] Kumar, S. Molecular clocks: four decades of evolution. *Nature*, 6(8):654–662, 2005.

[67] Pybus, O. G. and Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Publishing Group*, 10(8):540–550, 2009.

[68] Rambaut, A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4): 395–399, 2000.

[69] Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Andrew Rambaut. Relaxed Phylogenetics and Dating with Confidence. *Plos Biology*, 4(5):e88, 2006.

[70] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*, 29(8):1969–1973, 2012.

[71] Carroll, M. W., Matthews, D. A., Hiscox, J. A., et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature*, 524(7563):97–U201, 2015.

[72] Andersen, K. G., Shapiro, B. J., Matranga, C. B., et al. Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. *Cell*, 162(4):738–750, 2015.

[73] Frost, S. D. W., Pybus, O. G., Gog, J. R., et al. Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92, 2015.

[74] Wertheim, J. O., Fourment, M., and Kosakovsky Pond, S. L. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Molecular Biology and Evolution*, 29(2):451–456, 2012.

[75] Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1): 68–73, 1998.

[76] Huson, D. H. and Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–267, 2006.

[77] Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, 2014.

[78] Pinzon, J. E., Wilson, J. M., Tucker, C. J., et al. Trigger events: enviroclimatic coupling of Ebola hemorrhagic fever outbreaks. *The American journal of tropical medicine and hygiene*, 71(5): 664–674, 2004.

[79] Alexander, K. A., Sanderson, C. E., Marathe, M., et al. What factors might have led to the emergence of Ebola in West Africa? *PLOS Neglected Tropical Diseases*, 9(6):e0003652, 2015.

[80] Alizon, S., Lion, S., Murall, C. L., and Abbate, J. L. Quantifying the epidemic spread of Ebola virus (EBOV) in Sierra Leone using phylodynamics. *Virulence*, 5(8):825–827, 2014.

[81] Althaus, C. L. Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa. *PLoS Currents*, 6, 2014.

[82] Chowell, G. and Nishiura, H. Characterizing the transmission dynamics and control of ebola virus disease. *Plos Biology*, 13(1):e1002057, 2015.

[83] Chowell, G., Viboud, C., Hyman, J. M., and Simonsen, L. The Western Africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. *PLoS Currents*, 7, 2015.

[84] Fisman, D., Khoo, E., and Tuite, A. Early epidemic dynamics of the west african 2014 ebola outbreak: estimates derived with a simple two-parameter model. *PLoS Currents*, 6, 2014.

[85] House, T. Epidemiological dynamics of Ebola outbreaks. *eLife*, 3:e03908, 2014.

[86] Lewnard, J. A., Ndeffo Mbah, M. L., Alfaro-Murillo, J. A., et al. Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis. *Lancet Infect Dis*, 14(12):1189–1195, 2014.

[87] Meltzer, M. I., Atkins, C. Y., Santibanez, S., et al. Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014-2015. *MMWR. Morbidity and mortality weekly report.*, 63 Suppl 3:1–14, 2014.

[88] Nishiura, H. and Chowell, G. Early transmission dynamics of Ebola virus disease (EVD), West Africa, March to August 2014. *Eurosurveillance*, 19(36), 2014.

[89] Siettos, C., Anastassopoulou, C., Russo, L., Grigoras, C., and Mylonakis, E. Modeling the 2014 Ebola Virus Epidemic - Agent-Based Simulations, Temporal Analysis and Future Predictions for Liberia and Sierra Leone. *PLoS Currents*, 7, 2015.

[90] Towers, S., Patterson-Lomba, O., and Castillo-Chavez, C. Temporal variations in the effective reproduction number of the 2014 west Africa ebola outbreak. *PLoS Currents*, 6, 2014.

[91] World Health Organization. Ebola Situation Report - 20 January 2016, 2016. URL http://apps.who.int/ebola/current-situation/ebola-situation-report-20-january-2016.

[92] Ypma, R. J. F., van Ballegooijen, W. M., and Wallinga, Jacco. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–1062, 2013.

[93] Lemey, P., Derdelinckx, I., Rambaut, A., et al. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *Journal of Virology*, 79(18):11981–11989, 2005.

[94] Leitner, T., Escanilla, D., Franzén, C., Uhlén, M., and Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *PNAS*, 93(20):10864–10869, 1996.

[95] Paraskevis, D., Magiorkinis, E., Magiorkinis, G., et al. Phylogenetic reconstruction of a known HIV-1 CRF04_cpx transmission network using maximum likelihood and Bayesian methods. *Journal of Molecular Evolution*, 59(5):709–717, 2004.

[96] Jombart, T., Cori, A., Didelot, X., et al. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLOS Computational Biology*, 10(1):e1003457, 2014.

[97] Vega, V. B., Ruan, Y., Liu, J., et al. Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. *BMC Infectious Diseases*, 4:32, 2004.

[98] Morelli, M. J., Thébaud, G., Chadœuf, J., et al. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. *PLOS Computational Biology*, 8(11):e1002768, 2012.

[99] Didelot, X., Gardy, J., and Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol*, 31(7):1869–1879, 2014.

[100] Jombart, T., Aanensen, D. M., Baguelin, M., et al. OutbreakTools: A new platform for disease outbreak analysis using the R software. *Epidemics*, 7:28–34, 2014.

[101] Worby, C. J., Lipsitch, M., and Hanage, W. P. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLOS Computational Biology*, 10(3), 2014.

[102] Grenfell, B. T., Pybus, O. G., Gog, J. R., et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332, 2004.

[103] Lemey, P., Rambaut, A., and Pybus, O. G. HIV evolutionary dynamics within and among hosts. *AIDS reviews*, 8(3):125–140, 2006.

[104] Khiabanian, H., Carpenter, Z., Kugelman, J., et al. Viral diversity and clonal evolution from unphased genomic data. *BMC genomics*, 15 Suppl 6:S17, 2014.

[105] Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *PNAS*, 105(21): 7552–7557, 2008.

[106] Emmett, K. J., Lee, A., Khiabanian, H., and Rabadan, R. High-resolution Genomic Surveillance of 2014 Ebolavirus Using Shared Subclonal Variants. *PLoS Currents*, 7:–, 2015.

[107] Hall, M., Woolhouse, M., and Rambaut, A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology*, 11(12):e1004613, 2015.

[108] Worby, C. J., O'Neill, P. D., Kypraios, T., et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics*, 10(1): 395–417, 2016.

[109] Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., and Wallinga, Jacco. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5):e1005495, 2017.

[110] Cottam, E. M., Thébaud, G., Wadsworth, J., et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895, 2008.

[111] Mollentze, N., Nel, L. H., Townsend, S., et al. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B: Biological Sciences*, 281(1782), 2014.

[112] Numminen, E., Chewapreecha, C., Sirén, J., et al. Two-phase importance sampling for inference about transmission trees. *Proceedings of the Royal Society B: Biological Sciences*, 281 (1794), 2014.

[113] Worby, C. J., Chang, H.-H., Hanage, W. P., and Lipsitch, M. The distribution of pairwise genetic distances: a tool for investigating disease transmission. *Genetics*, 198(4):1395–1404, 2014.

[114] Campbell, F., Strang, C., Ferguson, N., Cori, A., and Jombart, T. When are pathogen genome sequences informative of transmission events? *PLoS pathogens*, 14(2):e1006885, 2018.

[115] Pybus, O. G., Charleston, M. A., Gupta, S., et al. The epidemic behavior of the hepatitis C virus. *Science*, 292(5525):2323–2325, 2001.

[116] Stadler, T., Kouyos, R., von Wyl, V., et al. Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, 29(1):347–357, 2012.

[117] Gomes, M. F. C., Pastore Y Piontti, A., Rossi, L., et al. Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLoS Currents*, 6, 2014.

[118] Lloyd-Smith, J. O., Cross, P. C., Briggs, C. J., et al. Should we expect population thresholds for wildlife disease? *Trends in Ecology & Evolution*, 20(9):511–519, 2005.

[119] Duffy, S., Shackelton, L. A., and Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature*, 9(4):267–276, 2008.

[120] Drake, J. W. Rates of spontaneous mutation among RNA viruses. *PNAS*, 90(9):4171–4175, 1993.

[121] Drake, J. W. and Hwang, C. B. C. On the mutation rate of herpes simplex virus type 1. *Genetics*, 170(2):969–970, 2005.

[122] Jenkins, G. M., Rambaut, A., Pybus, O. G., and Holmes, E. C. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*, 54(2):156–165, 2002.

[123] Pond, S. L. K. and Frost, S. D. W. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 21(10):2531–2533, 2005.

[124] Sealfon, R. S., Lin, M. F., Jungreis, I., et al. FRESCo: finding regions of excess synonymous constraint in diverse viruses. *Genome Biology*, 16:38, 2015.

[125] Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[126] Pond, S. L. K., Frost, S. D. W., and Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679, 2005.

[127] Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):275–276, 1977.

[128] Kryazhimskiy, S. and Plotkin, J. B. The population genetics of dN/dS. *PLoS Genetics*, 4(12): e1000304, 2008.

[129] Liu, S.-Q., Deng, C.-L., Yuan, Z.-M., Rayner, S., and Zhang, B. Identifying the pattern of molecular evolution for Zaire ebolavirus in the 2014 outbreak in West Africa. *Infection Genetics and Evolution*, 32:51–59, 2015.

[130] Wertheim, J. O. and Worobey, M. Relaxed selection and the evolution of RNA virus mucin-like pathogenicity factors. *Journal of Virology*, 83(9):4690–4694, 2009.

[131] Bedford, T., Cobey, S., and Pascual, M. Strength and tempo of selection revealed in viral gene genealogies. *BMC evolutionary biology*, 11:220, 2011.

[132] Neher, R. A., Russell, C. A., and Shraiman, B. I. Predicting evolution from the shape of genealogical trees. *eLife*, 3, 2014.

[133] Drummond, A. J. and Suchard, M. A. Fully Bayesian tests of neutrality using genealogical summary statistics. *BMC genetics*, 9:68, 2008.

[134] Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.

[135] Fu, Y. X. and Li, W. H. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, 1993.

[136] Kirkpatrick, M. and Slatkin, M. Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree. *Evolution*, 47(4):1171, 1993.

[137] McKenzie, A. and Steel, M. Distributions of cherries for two models of trees. *Mathematical biosciences*, 164(1):81–92, 2000.

[138] Colless, D. H. Review of: Phylogenetics: the theory and practice of phylogenetic systematics. *Syst Zool*, 31:100–104, 1982.

[139] Edwards, C. T. T., Holmes, E. C., Pybus, O. G., et al. Evolution of the Human Immunodeficiency Virus Envelope Gene Is Dominated by Purifying Selection. *Genetics*, 174(3):1441–1453, 2006.

[140] Lefever, S., Pattyn, F., Hellemans, J., and Vandesompele, J. Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. *Clinical chemistry*, 59 (10):1470–1480, 2013.

[141] Kugelman, J. R., Sanchez-Lockhart, M., Andersen, K. G., et al. Evaluation of the Potential Impact of Ebola Virus Genomic Drift on the Efficacy of Sequence-Based Candidate Therapeutics. *Mbio*, 6(1), 2015.

[142] Kugelman, J. R., Wiley, M. R., Mate, S., et al. Monitoring of Ebola Virus Makona Evolution through Establishment of Advanced Genomic Capability in Liberia. *Emerging Infectious Diseases*, 21(7):1135–1143, 2015.

[143] Castilletti, C., Carletti, F., Gruber, C. E. M., et al. Molecular Characterization of the First Ebola Virus Isolated in Italy, from a Health Care Worker Repatriated from Sierra Leone. *Genome announcements*, 3(3), 2015.

[144] Tong, Y.-G., Shi, W.-F., Liu, D., et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*, 524(7563):93–96, 2015.

[145] Simon-Loriere, E., Faye, O., Faye, O., et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*, 524(7563):102–U210, 2015.

[146] Diehl, W. E., Lin, A. E., Grubaugh, N. D., et al. Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic. *Cell*, 167(4):1088–1098.e6, 2016.

[147] Metcalf, C. J. E., Birger, R. B., Funk, S., et al. Five challenges in evolution and infectious diseases. *Epidemics*, 10:40–44, 2015.

[148] Yozwiak, N. L., Schaffner, S. F., and Sabeti, P. C. Data sharing: Make outbreak research open access. *Nature*, 518(7540):477–479, 2015.

[149] Simpson, J. T., Wong, K., Jackman, S. D., et al. ABySS: a parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123, 2009.

[150] Luo, R., Liu, B., Xie, Y., et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, 2012.

[151] Bankevich, A., Nurk, S., Antipov, D., et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5):455–477, 2012.

[152] Grabherr, M. G., Haas, B. J., Yassour, M., et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[153] Zerbino, D. R. and Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008.

[154] Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[155] Katoh, K. and Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, 2013.

[156] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[157] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[158] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[159] Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[160] Lee, W.-P., Stromberg, M. P., Ward, A., et al. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*, 9(3):e90581, 2014.

[161] Zaharia, M., Bolosky, W. J., Curtis, K., et al. Faster and More Accurate Sequence Alignment with SNAP. *arXiv*, 2011.

[162] Garrison, E. and Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv*, 2012.

[163] Rimmer, A., Phan, H., Mathieson, I., et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46 (8):912–918, 2014.

[164] Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res*, 21(3):487–493, 2011.

[165] Babraham Bioinformatics. FastQC, 2016. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[166] Rotmistrovsky, K. E. and Agarwala, R. BMTagger: Best Match Tagger for Removing Human Reads from Metagenomics Datasets, 2014. URL ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/.

[167] CLC bio. CLC Genomics Workbench. URL https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/.

[168] Rambaut, A. FigTree. URL http://tree.bio.ed.ac.uk/software/figtree/.

[169] Biomatters Ltd. Geneious. URL http://www.geneious.com/.

[170] Neher, R. and Bedford, T. Nextflu: Real-Time Tracking of Seasonal Influenza Virus Evolution in Humans. *Bioinformatics*, 31:3546–3548, 2015.

[171] Novocraft Technologies. NovoAlign, 2014. URL http://www.novocraft.com/products/novoalign/.

[172] Broad Institute. Picard. URL http://broadinstitute.github.io/picard/.

[173] Tomkins-Tinch, C., Ye, S., Metsky, H., et al. viral-ngs, 2016. URL doi:10.5281/zenodo.200428.

Chapter 2

Ebola Virus Epidemiology, Transmission, and Evolution During Seven Months in Sierra Leone

Preface

As noted in the previous chapter, the 2014–2016 Ebola outbreak in Western Africa was one of the first in which near-real-time whole-genome sequencing was used to understand a viral outbreak. Ebola virus (EBOV) first emerged in Guinea, West Africa in early 2014, having previously been observed only in Central Africa [1]. The virus then spread to Liberia, Sierra Leone, and Nigeria, ultimately resulting in over 28,000 cases and over 11,000 deaths [2]. Throughout the outbreak, a number of groups sequenced EBOV from patient samples [1, 3–11], totaling over 1,600 published EBOV genomes that could be used to analyze spread of the outbreak throughout Western Africa [12].

In Gire et al. [3], we released 99 EBOV genomes from 78 patients in Sierra Leone during the outbreak and identified three major phylogenetic clades (SL1, SL2, SL3) circulating within the country. As an author on the publication, I helped analyze these genomes, looking for patterns in the

sequences that could provide clues in understanding transmission during the outbreak. I also analyzed within-host variants identified in these EBOV sequences and used them to assess sequencing methods used in the study. In this chapter, based on Park D.J.*, Dudas, G.*, Wohl, S.*, Goba, A.*, Whitmer, S.L.M.*, et al. *Cell*, 2015 [5], I describe our follow-up study of EBOV in Sierra Leone, using an additional 232 EBOV genomes collected over seven months. I am a co-first author on this publication and have included the entire text of this collaborative work in this chapter for context, but I have highlighted my specific contributions below.

My primary contributions are detailed in Section 2.3.3, 'Human-to-Human Transmission of Multiple EBOV Genomes.' I performed the majority of the intrahost variant (iSNV) analysis, aiming to understand the patterns and prevalence of iSNVs in our dataset (Figures 2.2A, B.3A) and their relationship to sample coverage (Figure B.3B). As described below, these types of analyses helped rule out superinfection and contamination, and led to a more detailed understanding of EBOV transmission (Figures B.3C, 2.2B; the latter figure was generated by Trevor Bedford).

In collaboration with Danny Park and Gytis Dudas, I also explored variation at multiple timescales (Section 2.3.4). After noting the higher fraction of nonsynonymous variants in iSNVs as compared to consensus-level variants (Figure 2.3C), we explored the implications of this result on our purification selection hypothesis. I also looked for other evidence for selection or host effects on the viral genome (Section 2.3.5) in collaboration with Gytis Dudas and Aaron Lin. A preliminary d_N/d_S analysis highlighted the number of nonsynonymous variants within the EBOV glycoprotein, and specifically within the disordered mucin-like region. Gytis Dudas performed the more rigorous analysis shown in Figure 2.4A, and Aaron Lin led our efforts to investigate the conspicuous number

of T-to-C mutations throughout the genome (Figure 2.4B–F).

2.1 Abstract

The 2014–2016 Ebola virus disease (EVD) epidemic was caused by the Makona variant of EBOV. Early in the epidemic, genome sequencing provided insights into virus evolution and transmission and offered important information for outbreak response. Here, we analyze sequences from 232 patients sampled over seven months in Sierra Leone, along with 86 previously released genomes from earlier in the epidemic. We confirm sustained human-to-human transmission within Sierra Leone and find no evidence of import or export of EBOV across national borders after its initial introduction. Using high-depth replicate sequencing, we observe both host-to-host transmission and recurrent emergence of intrahost genetic variants. We trace the increasing impact of purifying selection in suppressing the accumulation of nonsynonymous mutations over time. Finally, we note changes in the mucin-like domain of EBOV glycoprotein that merit further investigation. These findings clarify the movement of EBOV within the region and describe viral evolution during prolonged human-to-human transmission.

2.2 Introduction

The 2014–2016 Western African EVD epidemic, caused by the EBOV Makona variant [13], is the largest EVD outbreak to date, with 26,648 cases and 11,017 deaths documented as of 8 May 2017 [14]. The outbreak, first declared in March 2014 in Guinea and traced back to the end of 2013 [1], has also

devastated the neighboring countries of Sierra Leone and Liberia, with additional cases scattered across the globe. Never before has an EBOV variant been transmitted among humans for such a sustained period of time.

Published EBOV Makona genomes from clinical samples obtained early in the outbreak in Guinea (three patients) and Sierra Leone (78 patients) [1, 3] demonstrated that near-real-time sequencing could provide valuable information to researchers involved in the global outbreak response. Analysis of these genomes revealed that the outbreak likely originated from a single introduction into the human population in Guinea at the end of 2013 and was then sustained exclusively by human-to-human transmissions. Genomic sequencing further allowed the identification of numerous mutations emerging in the EBOV Makona genome over time. As a consequence, the evolutionary rate of the Makona variant over the time span of the early phase of the outbreak could be estimated and predictions made about the potential of this new EBOV variant to escape current candidate vaccines, therapeutics, and diagnostics [15].

While the insights gleaned from sequencing early in the outbreak informed public health efforts [16–18], the continued human-to-human spread of the virus raises questions about ongoing evolution and transmission of EBOV. Our laboratory teams in Sierra Leone, at Kenema (Kenema Government Hospital [KGH]) and at Bo (U.S. Centers for Disease Control and Prevention [CDC]), continued to perform active diagnosis and surveillance in Sierra Leone following our initial study [3]. After a six-month delay of sample shipment due to regulatory uncertainty about inactivation protocols, we again began to determine EBOV genome sequences. We have sequenced samples at high depth and with technical replicates to characterize genetic diversity of EBOV both within

(intrahost) and between (interhost) individuals. To support global outbreak termination efforts, we publicly released these genomes prior to publication as they were generated, starting with a first set of 45 sequences in December 2014 and continuing with regular releases of hundreds of sequences through May 2015.

Here, we provide an analysis of 232 new, coding-complete EBOV Makona genomes from Sierra Leone. We compared these genomes to 86 previously available genomes: 78 unique genomes from Sierra Leone [3], three genomes from Guinea [1], and five from healthcare workers infected in Sierra Leone and treated in Europe. We use this combined dataset obtained from 318 EVD patients during the height of the epidemic in Sierra Leone and Guinea to better understand EBOV transmission within Sierra Leone and between countries. In addition, we use it to understand viral population dynamics within individual hosts, the impact of natural selection, and the characteristics of the now hundreds of new mutations that have emerged over the longer course of the epidemic.

2.3 Results

2.3.1 232 New Ebola Virus Makona Genomes from Sierra Leone

We performed massively parallel genome sequencing on 673 samples from two EVD patient cohorts. The first cohort included 575 blood samples from 484 EVD patients confirmed by laboratory staff at KGH from 16 June through 28 September 2014. The second cohort included blood samples from 88 EVD patients from throughout Sierra Leone confirmed at Bo by CDC laboratory staff from 20 August 2014 through 10 January 2015. Samples from both EVD cohorts were sequenced using

previously described methods (see Methods, Section 2.5) [3, 19].

We implemented a new computational pipeline, viral-ngs, for viral genomic de novo assembly, intrahost variant calling, and genome analysis and annotation. This pipeline is available via open-source software [20] and utilizes a generalized workflow engine to run on a wide variety of computer hardware configurations [21]. Through a partnership with DNAnexus, this pipeline is also available in a secure cloud-compute environment to enable consistent analyses across laboratories with limited computational resources (see Methods, Section 2.5).

Using this pipeline, we successfully assembled 232 EBOV Makona coding-complete genomes (150 from KGH and 82 from the CDC cohort, spanning 16 June to 26 December 2014). Each assembled sequence was at least 18.5 kb in length, with a maximum of 6% ambiguous base calls per genome. The median assembly had 374× coverage, was 18.9 kb long, and had no ambiguous bases. Despite extensive sequencing, successful full-genome assembly was difficult to obtain from the KGH cohort (73% failed genome assemblies; 374× mean coverage), compared to a previous cohort from the same laboratory, described in Gire et al. [3] (11% failed genome assemblies; 2,000× mean coverage). The high assembly failure rate of the more recent KGH cohort is likely due to the mandatory in-country implementation of a new EBOV sample deactivation protocol and to long delays for sample shipments amidst the outbreak response (see Methods, Section 2.5). In contrast, only 7% of samples from the CDC cohort failed to assemble. However, these samples had been preselected for sequencing based on high EBOV titers, as estimated by qPCR. In addition, the CDC cohort samples were collected more recently, did not remain in lysis buffer for an extended period, and were subjected to a different sample deactivation protocol than the KGH cohort samples.

While we are continuing attempts to glean genomic information from compromised samples of the recent KGH cohort, important information may have been lost. In particular, samples from many EBOV-infected health-care workers at KGH, which could provide important insights into hospital-based transmissions, were compromised.

In combination with the 86 previously published EBOV Makona genomes [3], we analyzed a total of 318 genomes (see Methods, Section 2.5), all aligned against the earliest sampled Guinean genome (GenBank accession: KJ660346.2). In this set, we observed 464 single-nucleotide polymorphisms (SNPs; 125 nonsynonymous, 176 synonymous, and 163 noncoding). We also observed five single-base insertions and two double-base insertions in noncoding regions. We mapped all of the variants to primer-binding sites for known sequence-based diagnostics [15] and found no mutations in these sites that were present in more than one Sierra Leonean sample.

We constructed a second, independent genome library for each of 150 high-quality samples from the KGH cohort to reliably determine iSNVs at low frequencies [3]. We identified 247 iSNVs (25 insertion/deletions that were excluded from all analyses, 73 nonsynonymous, 71 synonymous, and 78 noncoding), including 21 iSNVs shared by multiple patients.

Nearly simultaneously, another 175 EBOV Makona genomes were published based on a cohort from Sierra Leone, mostly sampled from the area of Freetown in the Fall of 2014 [4]. Although these data were not included in our analyses, they are unlikely to significantly alter our primary findings (Figure B.1).

2.3.2 Limited Ebola Virus Exchange Across the Sierra Leonean Border

A previous study of EBOV Makona sequences elucidated viral transmission and evolution during the early stages of the outbreak in Sierra Leone [3] from late May to early June, 2014. The first reported EVD cases in Sierra Leone stemmed from two genetically distinct EBOV Makona lineages, believed to have been introduced from Guinea. One of these lineages (SL1) was more closely related to the then-available three Guinean genomes (two to five mutations) than the second lineage (SL2), which was characterized by four additional mutations. This finding suggested that SL2 had evolved from SL1 some months before it was observed in Sierra Leone. A third lineage (SL3), derived from SL2, emerged in mid-June 2014. SL3 differs from SL2 by a single mutation at position 10,218, first found as an intrahost variant (polymorphism within one individual) at a low frequency. SL3 became the most prevalent lineage in Sierra Leone during the first three weeks of the outbreak there, with SL1 disappearing soon after the appearance of SL3. The SL3-defining mutation is epidemiologically important, as it is the first commonly circulating mutation observed to arise within Sierra Leone's borders.

As the epidemic developed within Sierra Leone, the SL3 lineage continued to dominate the viral population within the country, with no evidence for additional imported EBOV lineages. In our dataset, 97% of the genomes carry the SL3 mutation and the remainder belong to SL2 (Figure 2.1A). These results link all Sierra Leonean EVD cases to the initial introduction of EBOV into Sierra Leone, and they provide further evidence that all EVD cases during this outbreak arose from human-to-human transmission rather than from further zoonotic introductions from the unknown

EBOV reservoir. This means that no newly imported viral diversity was detected after the initial introduction [3]; all newly sampled viruses likely descended from those sequenced in the initial weeks of the outbreak. The genetic similarity of these viruses suggests that importation from other countries was minimal, although we cannot definitively rule out a re-introduction from elsewhere for the SL2 viruses (3%) in our dataset.

Similarly, publicly available EBOV genomes from this outbreak can shed light on exportation of EBOV from Sierra Leone into other countries. All published genomes from elsewhere, including 26 from Liberia and four from Mali, lack the Sierra Leone–defining SL3 mutation (Figure 2.1B and Methods, Section 2.5). Given that 97% of Sierra Leonean EBOV sequences have the SL3 variant, extensive exportation would result in the spread of SL3 EBOV genomes, a spread that is not seen in the limited samples available to date. At least in Sierra Leone, and with the exception of events at the onset of the epidemic, transmission has likely been primarily within national borders (Figure B.2 and Methods, Section 2.5), rather than by free interchange with neighboring countries.

2.3.3 Human-to-Human Transmission of Multiple EBOV Genomes

iSNVs that appear during the course of the epidemic may provide valuable information about human-to-human transmission. In particular, shared iSNVs have been used to estimate the relative size of the transmission bottleneck [22] and to identify human-to-human transmission chains [3]. In the current dataset, which includes 85 samples with at least one iSNV (Figure B.3A), several iSNVs are shared among two or more patients, often spanning several months of the EVD epidemic (Figure 2.2A). The existence of shared iSNVs could be explained by patient infection from multiple

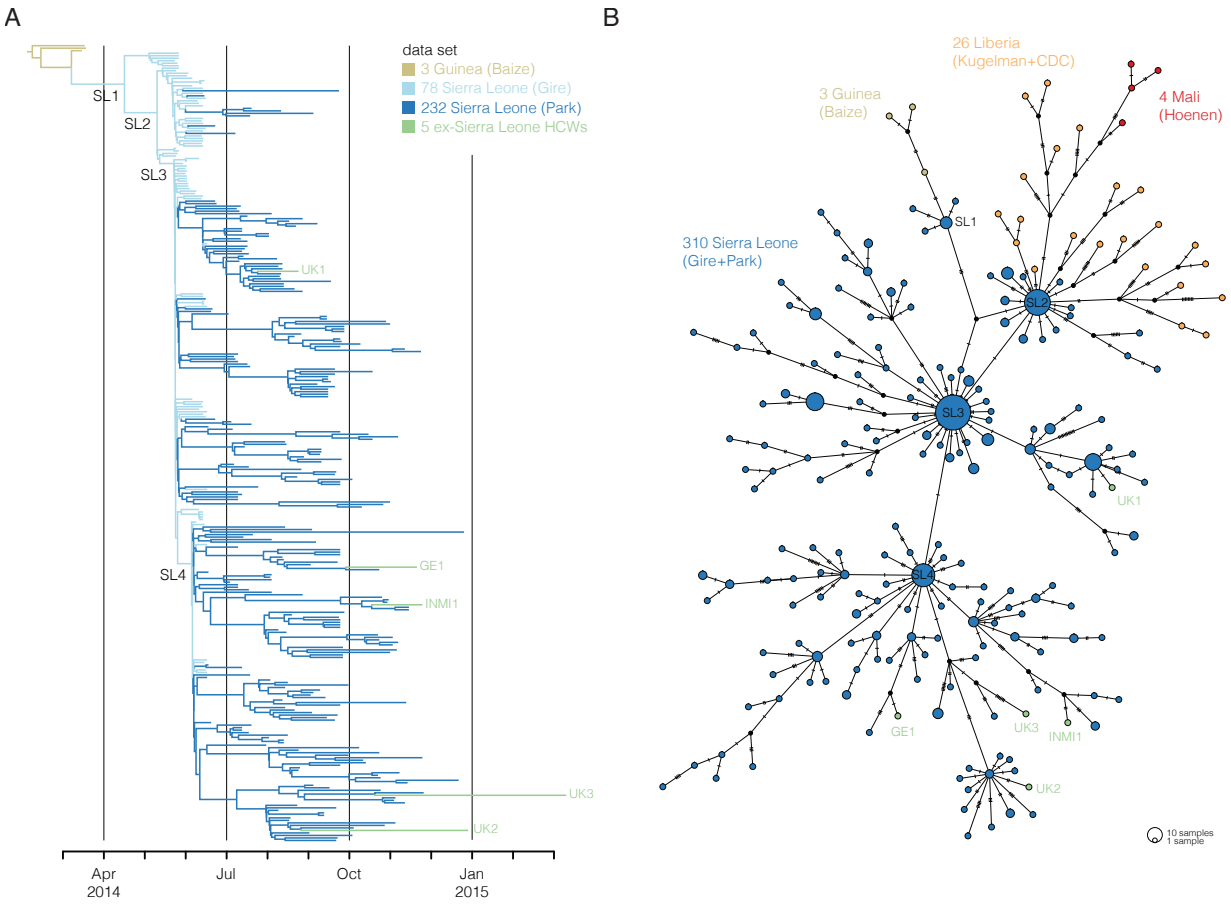


Figure 2.1: Within and between country genomic relationships of Ebola virus Makona. (**A**) Phylogenetic and temporal placement of recently sequenced EBOV within Sierra Leone. New EBOV genomes (232 genomes, dark blue), sampled from 16 June through 26 December 2014, provide a high-resolution view of the accumulated genetic diversity and fill in the missing ancestry between EBOV Makona genome datasets. The maximum clade credibility (MCC) tree was inferred using Bayesian evolutionary analysis by sampling trees (BEAST), with tips anchored to sampling date. Tips are labeled for EBOV from five non-African health-care workers (HCWs) infected in Sierra Leone and treated in Europe (sequenced by other groups, light green). Previously described nested EBOV Makona lineages SL1, SL2, and SL3 [3], as well as a new lineage SL4, are labeled at their most-recent common ancestor (MRCA) nodes. (**B**) Lack of EBOV Makona SL3 spread to Liberia or Mali. Shown is a median-joining haplotype network constructed from a coding-complete EBOV genome alignment including 340 EBOV Makona sequences. Each colored vertex represents a sampled viral haplotype, with colors indicating countries of origin. Colors are as in (A), with the exception that the distinction is no longer made between older (Gire) and newer (Park) Sierra Leonean datasets (both are now dark blue), and two additional countries are shown (Liberia in yellow, Mali in red). The size of the each vertex is relative to the number of sampled isolates. Hatch marks indicate the number of mutations along each edge. See also Figures B.1 and B.2.

sources (superinfection), sample contamination, recurring mutations (with or without balancing selection to reinforce mutations), or co-transmission of slightly diverged viruses that arose by mutation earlier in the transmission chain.

We can rule out superinfection and contamination as primary explanations for the iSNVs in our data because none of the iSNVs are located at common SNP positions. For example, a SNP at position 14,019 is at intermediate frequency in the population (found in approximately 40% of samples we sequenced) and defines the SL4 lineage (Figure 2.1A). If superinfection were common among EVD patients, we would expect to sometimes see both SL3 and SL4 viruses in the same patient, which would appear as an iSNV at that position. Contamination would result in a similar pattern, with intermediate-frequency SNPs appearing as iSNVs in contaminated samples. Additionally, contamination would be most visible in low-coverage, low-RNA-content samples because contaminants would make up more of the RNA available for sequencing, whereas samples with extremely high coverage would be the most visible contaminants (Figure B.3B). The highest coverage sample (G4960.1) contains genomes belonging to lineage SL3 only and lacks the SL4 SNP, so if there were widespread contamination, we would see a low-frequency iSNV at position 14,019 in SL4 samples with iSNVs. Since SL3 and SL4 samples were processed together (eight of nine sequencing batches contained multiple samples from both lineages) and we saw no instances of an iSNV at that position, we conclude that superinfection and contamination are not important contributors to iSNVs.

The remaining possible sources for persistently shared iSNVs are co-transmission and recurrent mutation. In either case, the iSNV could be maintained by balancing selection or could be evolving neutrally. Figure 2.2A suggests that selection is not the primary cause of persistence, since

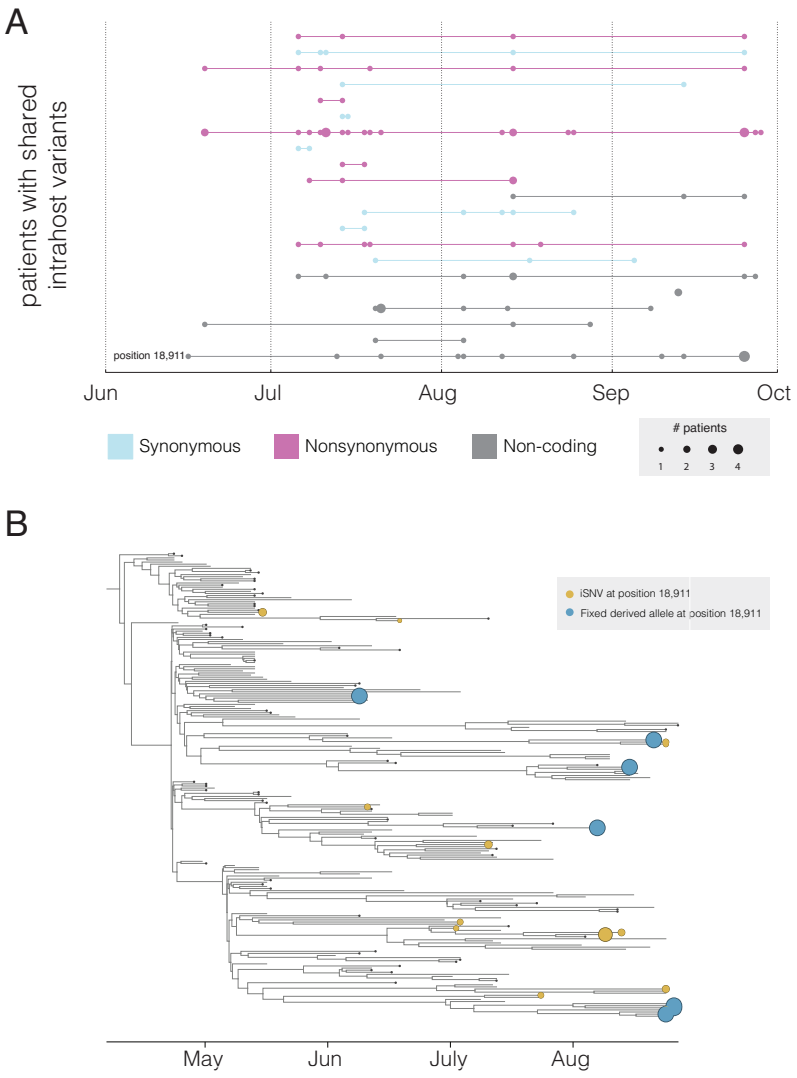


Figure 2.2: Evidence for host-to-host transmission of multiple Ebola virus Makona genomes. (**A**) Certain iSNVs appear in samples throughout the EVD epidemic, suggesting that iSNVs can be transmitted between patients. Variants shared between two or more samples are shown as rows of connected points; each row is a genomic position (ordered by position along the genome, top to bottom), and each point indicates the presence of the iSNV in a patient. (**B**) Phylogenetic placement of derived alleles at genomic position 18,911 implies both repeated transmission within clades as well as some amount of recurrent mutation. Colored tips are sized according to frequency of iSNV at position 18,911. Tips with small black points are those with iSNV calls at any position; other tips represent samples with no iSNV calls. This figure shows only the portion of the tree relevant for this analysis; large branches with no SNPs or iSNVs at position 18,911 are not shown. See also Figure B.3.

synonymous and nonsynonymous variants are equally common among the shared iSNVs, and selective pressures are likely to be different for the two classes of variant. All shared iSNVs are unlikely to be simply the product of recurring mutation: if they were, they should have a frequency spectrum heavily weighted toward low frequency, characteristic of new mutations. However, that is not the case. For example, the variant at position 18,911 is found at >15% frequency in eight different samples (Figure B.3C), a much higher frequency than expected if the change represented a de novo mutation in each sample.

In summary, we conclude that a combination of human-to-human transmission and recurrent mutations is likely responsible for the iSNV pattern observed in Figure 2.2A. This hypothesis is supported by the iSNV at position 18,911: samples containing this variant often cluster on the phylogenetic tree (Figure 2.2B), although more isolated samples may represent separate mutation events. More generally, pairs of samples that share an iSNV are typically located near one another phylogenetically; these pairs are separated by an average of 0.16 years of evolution, whereas random pairs are separated by an average of 0.30 years ($p < 10^{-4}$, randomization test). These results suggest transmission of iSNVs in at least some cases and therefore suggest that the transmission bottleneck is wide enough to facilitate the transmission of low- or intermediate-frequency variants between hosts.

2.3.4 Viral Evolution During a Prolonged EVD Epidemic

We previously reported that new mutations accumulated more rapidly in the viral population early in the outbreak than over the long-term in the reservoir [3]. We hypothesized then that the higher

rate early in the outbreak resulted from incomplete purifying selection — that is, we were detecting transient nonsynonymous variants that would later be removed by purifying selection [23, 24]. The observed evolutionary rate is thus not an estimate of the underlying mutation rate since some deleterious mutations are purged by selection before they can be detected. But neither is it an estimate of the long-term substitution rate since other deleterious mutations have not been eliminated by selection at the time of analysis. We hypothesized that the EBOV Makona evolutionary rate would decline following the addition of genomes covering a longer evolutionary timescale. Such a decline is well characterized in members of other species [25, 26]. With the present dataset, we were able to examine the evolution of the virus over a longer time period. We found that the most probable estimated evolutionary rate of EBOV Makona is indeed markedly lower (mean posterior rate = 1.25×10^{-3} substitutions per site per year) and is closer to the long-term rate than to the rate estimated early in the outbreak (Figures 2.3A and B.4).

How purifying selection acts at different timescales can also be seen in the distribution of mutations in the EBOV Makona genealogy. Deleterious mutations are more likely to result in transmission-impaired viruses and dead-end infections and may therefore only be present in individual patients. Mutations unique to individual patients are those that occur on the external branches of the phylogenetic tree, whereas internal branch mutations are those present in multiple samples in our dataset. Thus, in the model of incomplete purifying selection, we expect external branches to be characterized by a higher rate of nonsynonymous substitution than internal branches; in the latter, selection has had more opportunity to filter out deleterious mutants. Internal branches, by definition, have produced multiple descendent lineages and are thus less likely to

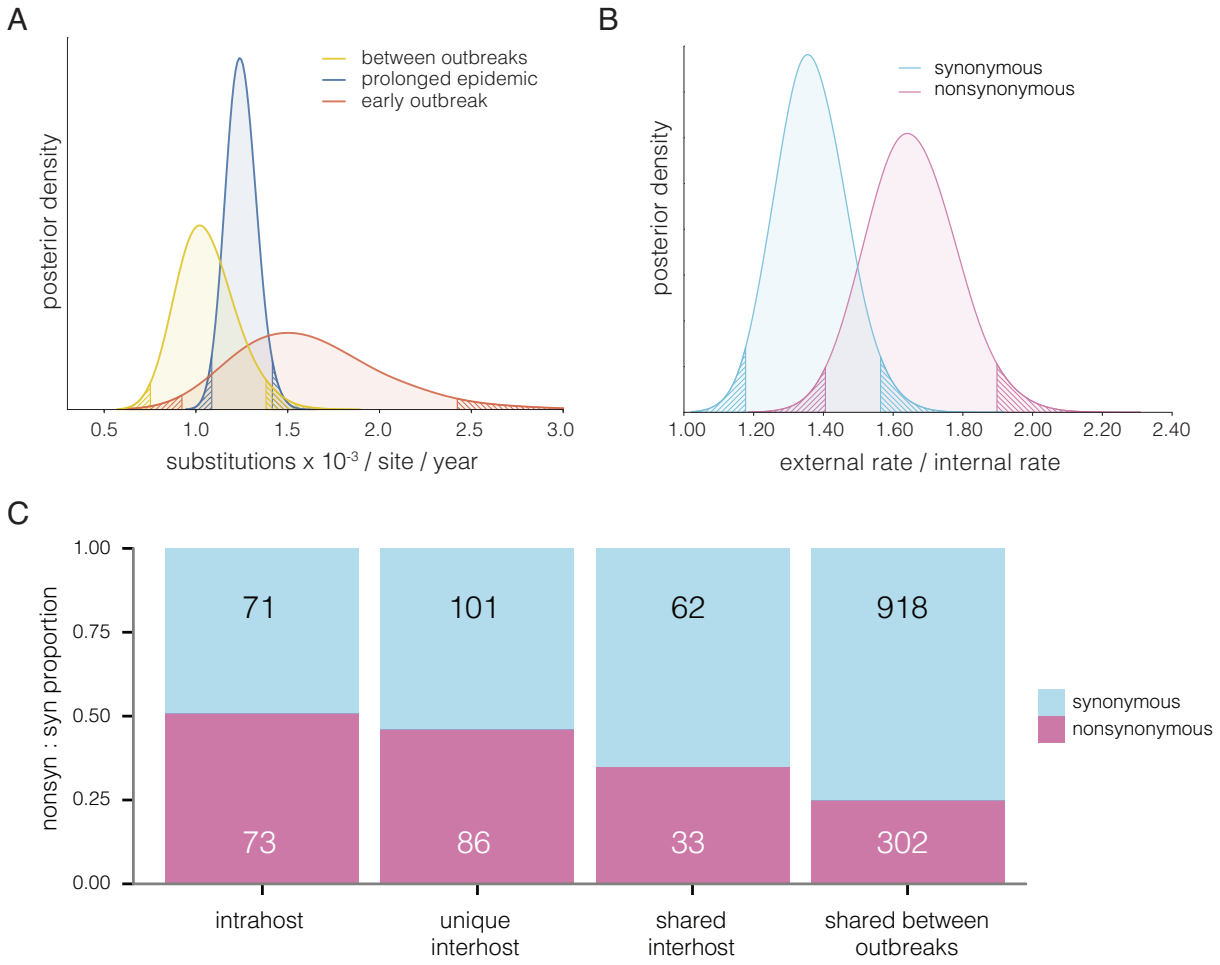


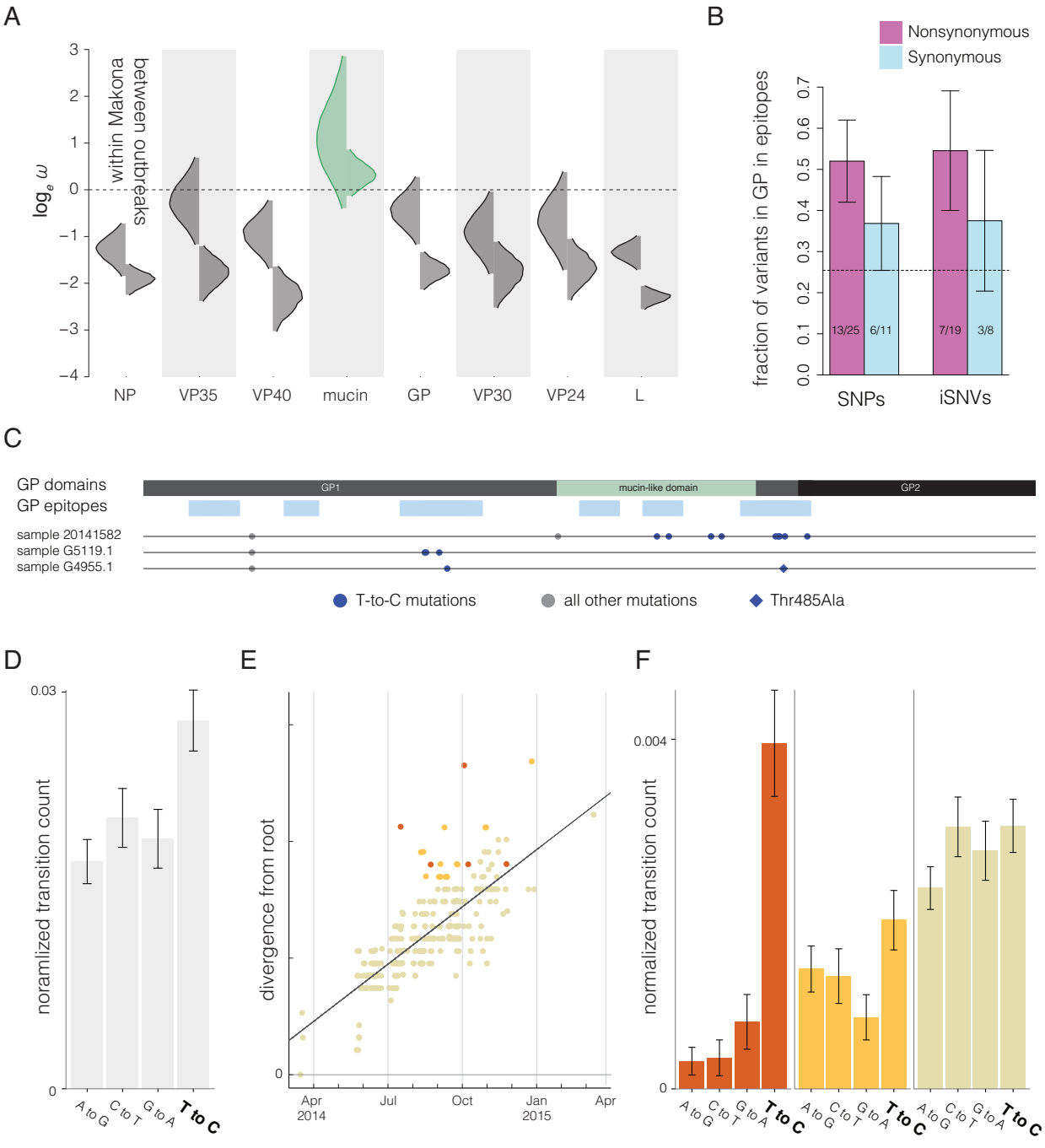
Figure 2.3: Ebola virus evolution during a prolonged EVD epidemic. (**A**) Estimates of EBOV evolutionary rates at three timescales: decades (yellow, all known EVD outbreaks), months (blue, Baize + Gire + Park), and weeks (red: Baize + Gire). (**B**) Purifying selection. We estimated nonsynonymous (red) and synonymous (blue) substitution rates on external (unique to an isolate, potential dead end) and internal (shared by multiple isolates, evidence of human-to-human transmission) branches. Nonsynonymous mutations accumulate faster on external branches than on internal branches. For synonymous mutations, the difference between external and internal branches is less pronounced. (**C**) Enrichment for nonsynonymous mutations at shorter timescales. Intrahost (all variants that appear within a single host at less than 100% frequency); unique interhost (SNPs fixed in exactly one individual); shared interhost (SNPs fixed in two or more individuals); shared between EVD outbreaks (internal branch SNPs on a between-outbreak tree). See also Figure B.4.

include mutations with fitness costs. To test this hypothesis, we estimated the numbers of nonsynonymous and synonymous changes on the virus genealogy and recovered their accumulation rates (Figure 2.3B). Nonsynonymous mutations indeed occurred at lower frequency on internal than on external branches, suggesting that most are removed by purifying selection because of their fitness costs and hence represent evolutionary dead ends. Synonymous mutations, which likely have less impact on fitness, occurred at more comparable frequencies on internal and external branches.

The relationship between the effectiveness of purifying selection and its duration is also apparent in the overall pattern of nonsynonymous mutations in our dataset. Selection filters the accumulation of coding variants in the EBOV genome (Figures 2.3C and 2.4A). Nonsynonymous mutations, which are more likely to be deleterious, make up a decreasing fraction of coding mutations as we analyze longer timescales: intrahost variants $>$ individual patients (external branches) $>$ multiple patients (internal branches) $>$ between outbreaks. The fraction seen between outbreaks represents the effect of long periods of evolution in the unknown EBOV reservoir. As selection acts to remove deleterious alleles over time, fewer nonsynonymous mutations can be detected. This pattern holds true across the EBOV Makona genome (Figure 2.4A).

Figure 2.4: Evidence for host effects on Ebola virus Makona evolution. (**A**) Nonsynonymous variants are enriched in the mucin-like domain of GP. Estimates of log(ω) (a.k.a., $log_e(d_N/d_S)$) per coding sequence within the Western African EVD outbreak (left) and between EVD outbreaks (right) demonstrate gene-specific patterns of natural selection. (**B**) Nonsynonymous variants are enriched in B cell epitopes of GP. We calculated the fractions of nonsynonymous (NS) and synonymous (S) consensus SNPs and iSNVs within experimentally determined B cell epitopes (data from ViPR [27]). Dotted line represents the fraction of GP amino acids in ViPR epitopes. Nonsynonymous SNPs (p = 0.004) and iSNVs (p = 0.037) in GP occur more frequently in epitopes than expected by chance (two-sided exact binomial test). Numbers indicate fraction of each variant type within GP epitope regions. Error bars represent binomial sampling intervals. (**C**) Local enrichment of T-to-C mutations within GP B cell epitopes. We observed five sequences with short stretches (<200 nucleotides) of concentrated T-to-C mutations. Of these five sequences, two (shown here, samples 20141582 and G5119.1) contain stretches of T-to-C SNPs (blue points) within GP epitopes (light blue bars). Additionally, we observe a T-to-C mutation at amino acid position 485 (blue diamond) in three samples (one shown here, G4955.1), which is otherwise completely conserved among members of all ebolavirus species [28]. (**D**) Genome-wide increase in T-to-C mutations. We observe more T-to-C transitions within the 2014–2016 outbreak than any other transition, after correcting for nucleotide content. Error bars represent binomial sampling intervals. (**E-F**) Elevated T-to-C rates are genome wide but are limited to a subset of sequences. Accumulation of mutation increases linearly with time. However, some individual samples show more genetic distance than expected based on sample date. Samples with short stretches of T-to-C mutations (orange) show a significant enrichment of T-to-C mutations, as expected. Excluding these samples, the top 5% of samples by genetic distance (yellow) lack localized stretches but still show moderate enrichment of T-to-C mutations genome wide. The bottom 95% of samples (beige) show no enrichment of T-to-C mutations. Error bars represent binomial sampling intervals.

Figure 2.4: Continued



2.3.5 Possible Host Effects on the Viral Genome

Although we observe less constraint on nonsynonymous changes during the 2014–2016 epidemic than between outbreaks, one anomaly is the genomic sequence encoding the mucin-like domain of the EBOV glycoprotein (GP), for which we observe more nonsynonymous substitutions than expected under neutrality, both within and between EVD outbreaks. Selective pressure acting on a region can be estimated with the standard statistic d_N/d_S, which has an expected value of 1.0 for neutral evolution and less than 1.0 for purifying selection; in the mucin-like domain, the mean posterior d_N/d_S within this outbreak is 4.74, and between outbreaks is 1.44 (Figure 2.4A). GP is the only surface-exposed viral protein on EBOV virions, and as such, it is the primary target of antibodies [29]. This finding therefore raises the possibility that antibodies might be driving diversifying selection and rapid evolution in this region. This observation is based on a very small number of substitutions (eight nonsynonymous and four synonymous within the outbreak), however, and is not statistically significant (posterior probability that d_N/d_S is elevated within-outbreak = 92.9%); the situation should be clarified as more sequencing becomes available. If diversifying selection is occurring here, then the observed changes are very unlikely to represent population-level selection for transmission among humans; this would only occur if previously infected individuals were frequently being exposed to new infections. Instead, we hypothesize that these changes represent within-host selection for EBOV to escape a developing humoral immune response.

To test the hypothesis that antibodies drive diversifying selection of GP, we looked for enrichment of mutations within B cell epitopes within that protein. Effective humoral immunity depends

on antibody binding to specific B cell epitopes [29, 30]. Using experimentally determined B cell epitopes obtained from the Virus Pathogen Database and Analysis Resource (ViPR) [27], we found that nonsynonymous mutations in GP do indeed occur more frequently in epitopes than expected by chance (Figure 2.4B). This correlation supports the hypothesis that humoral immunity exerts selective pressure on the virus, driving immune evasion via accumulation of nonsynonymous mutations within GP B cell epitopes.

Visual inspection identified a subset of sequences that are more likely to contain B cell escape variants (Figure 2.4C). In particular, three sequences (e.g., G4955.1) had a threonine-to-alanine mutation at GP amino acid position 485, a conserved threonine that is required for in vivo protection by the 14G7 antibody [28]. Additionally, two sequences had short stretches of T-to-C mutations in GP (four or more T-to-C mutations within a 200 nucleotide region; Figure 2.4C), both of which occur within B cell epitopes.

Similar patterns of excess T-to-C mutations within short regions were also observed by Tong et al. [4]. In our dataset of 318 genomes, five possessed obvious stretches of T-to-C mutations within short regions. We also tested more broadly whether excessive T-to-C mutations occurred in all sequences and found a significant enrichment of T-to-C transitions relative to all other types of transitions (Figure 2.4D). To determine whether viral sequence divergence is related to T-to-C transition enrichment, we compared relative T-to-C transition rates in sequences with stretches of T-to-C mutations (n = 5) to the top 5% of remaining sequences by sequence divergence (n = 15) and to the bottom 95% of sequences (n = 298) (Figure 2.4E). While the sequences with T-to-C stretches showed the strongest T-to-C enrichment, we found moderate enrichment of T-to-C transitions in

the 5% most divergent sequences.

2.4 Discussion

Our findings from 232 EBOV Makona genomes sampled in Sierra Leone over seven months during the 2014–2016 EVD outbreak in Western Africa demonstrate the value of continued sequencing throughout an epidemic. We tracked the movement of EBOV throughout Sierra Leone and determined the frequency of EBOV movement into and out of that country. Although it is not unlikely that the virus continued to cross the national borders of Sierra Leone throughout the epidemic, these observations suggest that, at least in late 2014, cross-border introductions were not an important factor in the development of the epidemic. We were unable, however, to draw any conclusions about export to Guinea since few EBOV sequences from there were available at the time.

The sequence data display EBOV Makona evolution in the context of prolonged human-to-human transmission and provide an updated view of genomic diversity. Based on the rates of non-synonymous and synonymous changes that are shared or are unique to an individual host, we concluded that purifying selection becomes increasingly effective over time, as it has more opportunity to remove deleterious mutants.

While the effects of purifying selection in this extended EVD outbreak are clear, these evolutionary changes do not imply that positive selection or adaptation to humans are occurring. Rather, the data suggest that evolutionary changes over time through natural selection are sufficient to remove newly arisen alleles that are less fit in the human environment.

It is important to recognize, however, that the long-term human-to-human transmission observed during the 2014–2016 EVD outbreak is historically unique for EBOV. At the beginning of each EVD outbreak, EBOV enters the human population with little or no genetic diversity. In the case of the current EVD outbreak, EBOV has now maintained fitness while expanding across a much larger space of genetic diversity than in previous EVD outbreaks, the largest of which comprised only 318 human infections. This degree of diversity will undoubtedly affect researchers' ongoing efforts to develop or improve candidate diagnostics, vaccines, and therapeutics for EVD, many of which are targeting EBOV sequences directly (PCR, nucleic-acid based therapeutics) or indirectly (antibody cocktails).

The mucin-like domain of the EBOV glycoprotein, in contrast to the rest of the EBOV genome, appeared to be under diversifying selection based on a high ratio of nonsynonymous-to-synonymous mutations. While not statistically significant because of the small number of SNPs in the region, our observation is in agreement with many previous studies [31, 32]. As the EBOV GP, especially the mucin-like domain, is the target of many antibodies, a plausible hypothesis is that the humoral immune response exerts selective pressure on GP, resulting in an accumulation of nonsynonymous mutations. In support of this hypothesis, regions of GP corresponding to experimentally determined B cell epitopes are significantly enriched in nonsynonymous, but not in synonymous, variants. There are two important caveats to this analysis: (1) these epitopes are determined in vitro and therefore may not be epitopes in vivo if they are not immunodominant, and (2) there is no experimental evidence to suggest that the majority of observed variants disrupt antibody binding to these epitopes.

While further experimental testing is required to validate an immune evasion hypothesis, we have highlighted a few prime candidates to consider. Genomes from three samples share a threonine-to-alanine mutation at GP amino acid position 485, a position that is conserved among all members of the Ebolavirus genus. This position is indispensable for binding of the protective antibody 14G7 [28]; the observed variant at this site may therefore be the result of escape from antibody-mediated selection. Additionally, two samples each possess multiple mutations within a single experimental B cell epitope in GP, which are likely to evade antibody recognition if those regions are relevant epitopes in vivo.

Intriguingly, the two samples with multiple mutations within a single B cell epitope each possess a distinct short stretch littered with T-to-C transitions, a phenomenon also observed in Tong et al. [4]. Excessive T-to-C and A-to-G mutation of virus genomes has been observed previously as a result of adenosine deaminases acting on RNA (ADARs) [33–35]. When acting on viral genomic RNA, ADARs cause a pattern of excess A-to-G transitions that are represented by T-to-C transitions in our dataset. These transitions are known to occur either promiscuously within 200 nucleotide stretches or in a sequence-specific manner; therefore, we investigated both possibilities. While only five of the 318 sequences in our dataset contained obvious T-to-C stretches, we showed that the top 5% of sequences by sequence divergence, excluding the five sequences with T-to-C stretches, were also moderately enriched for T-to-C transitions across the genome. The remaining 95% of sequences appeared to show no enrichment. We do not know whether this phenomenon is caused by ADAR acting upon genomic RNA, as we cannot exclude the possibility of bias by the EBOV RNA polymerase or other effects. Additionally, it is yet unclear whether these T-to-C muta-

tions have an anti-viral or other effect on viral fitness. These questions open avenues of research into molecular mechanisms shaping EBOV evolution.

The results of some of the specific genome analysis methods that we introduced here, while promising, will require denser EBOV genome sampling to yield sufficient information to influence the EVD outbreak response. Among these methods is transmission analysis, which could prove valuable for improved understanding of hospital-based transmissions and therefore for improved infection control. Inference of the ancestral genetic state is often straightforward, with clear patterns of new variations layering on previously existing variations; viruses that appear to be descended from others in the same dataset are separated only by new mutations that are seen nowhere else in the dataset. This kind of genetic relationship does not guarantee a transmission relationship between two patients since many viruses can share identical genomes. However, since viruses with identical genomes are often epidemiologically related [3], we can infer that viruses that appear to descend from other viruses in our dataset are either in or epidemiologically close to the same transmission chain.

Unfortunately, long delays of shipping samples from the field and required changes to the EBOV inactivation protocol caused severe degradation of many samples, which prevented identification of variants and transmission analysis. This loss should serve as a reminder that standardized and optimized protocols for sample collection, virus deactivation, and shipment are crucial for a rapid worldwide response to any new infectious disease outbreak. An important future research effort will be aimed at understanding which certified EVD sample deactivation protocols are best suited for high-quality genomic sequencing. Complications with sample shipment also emphasize

the need for establishing in-country sequencing capabilities either before or at the onset of future EVD outbreaks [36].

Beyond coordinated field and experimental responses, a culture of rapid data sharing is critical for teams around the world to have the best current information about a circulating virus or ongoing disease [37]. In light of this need, we released all data discussed in this paper publicly as they were generated, beginning in December 2014, well in advance of our own analysis. We have previously described our high-depth sequencing protocols [19], and we have also made available our computational analysis pipeline, in the hope that they will assist the many laboratories engaged in viral genomic research. More EBOV data will allow the scientific community to together obtain a broader picture of transmission and evolution of EBOV Makona during the EVD epidemic.

2.5 Methods

2.5.1 Ethical and Safety Approvals

This work has been approved by Institutional Review Boards in Sierra Leone (Sierra Leone Ethics and Scientific Review Committee, SLESRC) and the United States (Harvard Committee on the Use of Human Subjects, CUHS, the CDC's Human Research Protection Office, HSPO). As part of the EVD outbreak response and surveillance efforts, residual human clinical samples were collected under a waiver of consent granted by SLESRC and CUHS, and the EBOV sequencing work has received non-human subjects research determination by CUHS and HSPO. The Sierra Leone Ministry of Health and Sanitation approved shipment of non-infectious, inactivated samples collected

from EVD patients to Broad Institute and Harvard University for viral sequencing. The EBOV-related research and laboratory safety protocols are registered with the Committee of Microbiological Safety (COMS) at Harvard University, and the viral sequencing work is registered with the Institutional Biosafety Committee at Broad Institute. All work with infectious or potentially infectious material was performed at the CDC Viral Special Pathogens Branch in Atlanta, GA, under biosafety level 4 (BSL-4) conditions. Our work was not deemed to be dual-use research of concern.

2.5.2 Sample Preparation from Kenema Government Hospital

This study included 575 blood samples from 84 patients with confirmed EVD from 16 June through 28 September 2014 by KGH laboratory staff. Clinical samples were inactivated using QIAGEN AVL and ethanol in the KGH laboratory prior to shipping out of the country.

2.5.3 Sample Preparation from CDC Bo Laboratory

This study included 98 blood samples from 98 patients with confirmed EVD from 20 August 2014 through 10 January 2015 by CDC laboratory staff stationed in Bo, Sierra Leone. Clinical specimens from the CDC Bo laboratory in Sierra Leone were shipped to and stored at the Viral Special Pathogens Branch BSL-4 laboratory at the CDC in Atlanta, GA. Samples were inactivated and RNA was extracted using the MagMAX Pathogen RNA/DNA isolation kit (Invitrogen) and BeadRetriever (Invitrogen). Non-infectious RNA was treated with DNase I RNase-free (Roche) prior to shipment to the Broad Institute.

2.5.4 High-Throughput Sequencing

Host ribosomal and carrier poly(rA) RNA depletion, randomly primed cDNA synthesis, Nextera XT library construction, and 101-bp paired-end Illumina sequencing were performed as described previously [3, 19].

2.5.5 Ebola Virus Makona Genome Assembly and Analysis

EBOV Makona genomes were assembled from high-throughput sequencing data using an updated bioinformatics pipeline based on our previously described methods [3, 19]. Of the collected samples, 150 KGH and 82 CDC samples had sufficient EBOV genome sequencing coverage for high-quality de novo genome assembly.

The viral assembly pipeline began by depleting paired-end reads from each sample of human and other contaminants using best match tagger (BMTagger) [38] and the nucleotide basic local alignment search tool (BLASTN) [39]. PCR duplicates were removed using a custom modification to Vicuna, M-Vicuna [40]. The resulting "de-identified" metagenomic datasets were deposited in sequence read archive (BioProjects PRJNA257197 and PRJNA283385). Next, reads were filtered to all members of the Ebolavirus genus (all ebolaviruses including EBOV) using LASTAL [41], quality-trimmed with Trimmomatic [42], and further de-duplicated with PRINSEQ [43].

The filtered and trimmed reads were subsampled to 100,000 pairs, if available, and de novo assembled using Trinity [44]. Subsequently, reference-assisted assembly improvements (contig scaffolding, gap-filling, etc.) were performed with virtual file application table [45], which relies

on MOSAIK [46] and multiple sequence comparison by log expectation (MUSCLE) [47]. Each sample's reads were aligned to its de novo assembly using Novoalign [48], and any remaining duplicates were removed using Picard with MarkDuplicates command [49]. Variant positions in each assembly were identified using genome analysis toolkit [50] insertions and deletions realinger (IndelRealigner) and UnifiedGenotyper [51, 52] on the read alignments. The assembly was refined to represent the major allele at each variant site, and any positions supported by fewer than three reads were changed to N (nonsynonymous sites). This align-call-refine cycle was iterated twice, to minimize reference bias in the assembly.

Our Linux-based software pipeline is publicly available at `https://github.com/broadinstitute/viral-ngs` [20]. This pipeline includes command-line tools for each of the above steps and optional Snakemake workflows [21] to automate them either sequentially or in parallel. Most of the third-party tools used are either included or can be downloaded and installed automatically, except for GATK and Novoalign, which must be provided by the user due to licensing restrictions.

The assembly pipeline is also available via the DNAnexus cloud platform. RNA paired-end reads from either HiSeq or MiSeq instruments (Illumina) can be securely uploaded in FASTQ or BAM format and processed through the pipeline using graphical and command-line interfaces. Instructions for the cloud analysis pipeline are available at `https://github.com/dnanexus/viral-ngs/wiki`.

2.5.6 Genomic Epidemiology of Ebola Virus Makona

The following publicly available EBOV Makona genomes from outside of Sierra Leone do not carry the SL3-derived allele at position 10,218: 26 available genomes from Liberia (25 from [53], one from GenBank accession: KP178538.1), and all four available genomes from Mali [8]. A median-joining haplotype network was constructed in PopART version 1.7.2 (http://popart.otago.ac.nz). Due to the presence of missing data, 1,492 sites (7.9% of total genome) were excluded from the analysis; these sites included 61 sites with variability among isolates (10.9% of all variable sites).

To reconstruct the EBOV Makona transmission history within Sierra Leone, we grouped samples into sets of one or more genetically identical viruses based on their consensus sequences. We then identified relationships between these groups, progressing from the Guinean reference genome (KJ660346.2) and ending with nine viruses sampled in Freetown (eight from our KGH and CDC cohorts and one sequenced in Italy).

2.5.7 Intrahost Variant Analysis

iSNVs were called from each sample's read alignments using V-Phaser 2.0 [54] and subjected to an initial set of filters: variant calls with fewer than five forward or reverse reads or more than a 10-fold strand bias were eliminated. iSNVs were also removed if there was more than a five-fold difference between the strand bias of the variant call and the strand bias of the reference call. Variant calls that passed these filters were additionally subjected to a 0.5% frequency filter. The final list of iSNVs contains only variant calls that passed all filters in two separate library preparations. These

data infer 100% allele frequencies for all samples at an iSNV position without intrahost variation within the sample, but a clear consensus call during assembly. Annotations were computed with the effect of single nucleotide polymorphisms (SnpEff) program [55].

Evolutionary distances between pairs of phylogeny tips were computed from the posterior sample of trees produced by Bayesian evolutionary analysis by sampling trees (BEAST) [56] analysis. This calculation integrates across phylogenetic uncertainty and produces a temporal evolutionary distance between phylogeny tips. We used this distance matrix to calculate the average distance between pairs of phylogeny tips that share an iSNV and compared the result to the average distance between random pairs of tips. We calculated a *p* value for the observed average distance by conducting a randomization test. In each random replicate, we sampled the same distribution of iSNV possessing tips as observed in the empirical data and calculated the average distance between these pairs of tips. We calculated a *p* value by comparing the empirical mean distance to the mean distances observed over 10,000 random replicates.

2.5.8 GP B Cell Epitope Analysis

Data were obtained from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) online through the web site at `http://www.viprbrc.org` [27]. As most of the epitopes in the database are based on the Mayinga reference strain, we mapped all B cell epitopes against the Guinean reference strain (GenBank accesion: KJ660346.2) and removed all epitopes that no longer matched perfectly, leaving 40 B cell epitopes. Overlapping epitopes were merged, and nonsynonymous and synonymous SNPs and iSNVs were scored as within or outside of epitope regions. Significance

was determined by two-tailed binomial test with $\alpha = 0.05$, with the null hypothesis that variants would occur in epitope regions of GP by chance with probability 172/676, which is the fraction of GP residues GP within B cell epitopes.

2.5.9 Molecular Evolution

Three datasets were constructed to represent three timescales of genetic surveillance of EBOV Makona. For surveillance between EVD outbreaks, 63 publicly available sequences represent the diversity of EBOV sampled over long periods of time; these sequences include the first recorded 1976 EVD outbreak and other EVD outbreaks and exclude one outbreak occurring in the Democratic Republic of the Congo in 2014. We also included EBOV genome fragment sequences from possibly infected great ape carcasses and frugivorous bats. Fourteen sequences from Western Africa were chosen to represent the 2014–2016 EVD outbreak. For surveillance of the early outbreak, 81 sequences [1, 3] were reanalyzed, representing the earliest epidemiologically relevant and publicly available sequences. For surveillance of the prolonged epidemic, 232 EBOV genomes reported here were combined with five sequences from repatriated healthcare workers (UK1, UK2, UK3, INMI1, GE1) and the 81 sequences from the early outbreak dataset.

Analyses of rates, phylogenies, and evolution were performed on all three datasets in BEAST [56]. Synonymous and nonsynonymous counts were mapped onto the molecular phylogenies using robust counting [57, 58] by specifying independent Hasegawa, Kishno, Yano (HKY) nucleotide substitution models [59] for all three codon-position partitions. Substitutions in intergenic regions were modeled according to HKY with $\Gamma 4$-distributed rate heterogeneity [59, 60]. A relaxed molec-

ular clock with log-normal rate distribution categories [61] and a non-parametric Bayesian skygrid [62] tree prior were used. A reference prior [63] was used on the molecular clock.

We estimated the ratio of nonsynonymous substitutions over synonymous substitutions, d_N/d_S or ω in every gene of EBOV (NP, VP35, VP40, VP30, VP24 and L), using an implementation of the Goldman & Yang [64] codon model in BEAST. We used the same sequences as the analysis above, but excluded sequences of potentially lower quality, resulting in 314 EBOV Makona genomes. GP-gene coding sequences were split into the mucin-like domain ($GP1_{MLD}$), which encompasses amino acid residues 313–464 [65] starting from methionine of GP1, and the rest of GP1,2 ($GP_{\Delta MLD}$). This split is due due to concern that the GP_{MLD} is highly disorganized [66] and thus is under little constraint at the amino acid level. To date, only linear epitopes in GP_{MLD} are known to be targeted by antibodies [28], due to its extensive O- and N-linked glycosylation. We employed independent codon models for all eight partitions, parameterized with independent strict molecular clocks. A reference prior [63] was used on the evolutionary rate. Substitutions in the ninth partition, with concatenated noncoding intergenic regions, was modeled using the $HKY + \Gamma 4$ [59, 60] model. The non-parametric Bayesian skygrid was used as the tree prior [62] for both long-term and current datasets.

2.6 Acknowledgements

We thank KGH staff who died of EVD (including M. Fonnie, A. Moigboi, A. Kovoma, M. Fullah, and S.H. Khan), the Office of the President of Sierra Leone (President E. Koroma, M. Jones), the Sierra

Leone Ministry of Health and Sanitation, the Kenema District Health Management team, and the Kenema Lassa fever program for their immense efforts in the EVD outbreak response. We thank Public Health England (UK1, UK2, UK3), IRCCS Lazzaro Spallanzani (INMI1), and the University of Geneva (GE1) for providing EBOV genome sequences from samples of EVD patients exported from Sierra Leone. We thank the drivers, pilots, phlebotomists, non-governmental organizations, district medical officers, and district surveillance officers for their help with sample collection and logistics in Sierra Leone. We want to especially thank the Médecins Sans Frontières (MSF) operation centers for their continuing support of the U.S. Centers for Disease Control and Prevention (CDC) laboratory in Bo, Sierra Leone and the World Health Organization (WHO) for their support of the preceding CDC laboratory operation in Kenema, Sierra Leone.

This work was supported by European Union grant FP7/2007-2013 278433-PREDEMICS and European Research Council grant 260864 (A.R.); Natural Environment Research Council grant D76739X (G.D.); NIH U54 GM111274 (T.B.); NIH grant GM080177 (S.W.); NIH grant 1U01HG007480-01 (C.H.); National Science Foundation Graduate Research Fellowship Grant No. DGE 1144152 (A.E.L.); the National Health and Medical Research Council, Australia (E.C.H.); the Defense Threat Reduction Agency (USAMRIID); NIH/NIAID U19AI110818 (Broad Institute); the Bill and Melinda Gates Foundation OPP1123407 (Broad Institute); and NIAID HHSN272200900049C (Harvard/Tulane). This work was funded, in part, through Battelle Memorial Institute's prime contract with the U.S. National Institute of Allergy and Infectious Diseases (NIAID) under contract number HHSN272200700016I. Subcontractors to Battelle Memorial Institute who performed this work are: J.H.K., an employee of Tunnell Government Services, Inc. R.F.G. is co-founder of Zalgen Labs.

The Virus Pathogen Database and Analysis Resource (ViPR) has been wholly funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract number HHSN272201400028C.

The content of this chapter does not necessarily reflect the views or policies of the U.S. Department of Health and Human Services (Centers for Disease Control and Prevention, National Institutes of Health) or the U.S. Army.

2.7 References

[1] Baize, S., Pannetier, D., Oestereich, L., et al. Emergence of Zaire Ebola Virus Disease in Guinea — Preliminary Report. *New England Journal of Medicine*, 2014.

[2] Organization, W. H. Ebola Situation Reports - 10 June 2016, 2016. URL http://apps.who.int/iris/bitstream/10665/208883/1/ebolasitrep_10Jun2016_eng.pdf.

[3] Gire, S. K., Goba, A., Andersen, K. G., et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[4] Tong, Y.-G., Shi, W.-F., Liu, D., et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*, 524(7563):93–96, 2015.

[5] Park, D. J., Dudas, G., Wohl, S., et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526, 2015.

[6] Simon-Loriere, E., Faye, O., Faye, O., et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*, 524(7563):102–U210, 2015.

[7] Ladner, J. T., Wiley, M. R., Mate, S., et al. Evolution and Spread of Ebola Virus in Liberia, 2014-2015. *Cell Host and Microbe*, 18(6):659–669, 2015.

[8] Hoenen, T., Safronetz, D., Groseth, A., et al. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science*, 348(6230):117–119, 2015.

[9] Carroll, M. W., Matthews, D. A., Hiscox, J. A., et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature*, 524(7563):97–U201, 2015.

[10] Folarin, O. A., Ehichioya, D., Schaffner, S. F., et al. Ebola Virus Epidemiology and Evolution in Nigeria. *The Journal of Infectious Diseases*, 214(suppl 3):S102–S109, 2016.

[11] Quick, J., Loman, N. J., Duraffour, S., et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–233, 2016.

[12] Dudas, G., Carvalho, L. M., Bedford, T., et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, 544(7650):309–315, 2017.

[13] Kuhn, J. H., Andersen, K. G., Baize, S., et al. Nomenclature- and database-compatible names for the two Ebola virus variants that emerged in Guinea and the Democratic Republic of the Congo in 2014. *Viruses*, 6(11):4760–4799, 2014.

[14] Organization, W. H. Ebola Situation Reports, 2015. URL http://apps.who.int/ebola/en/ebola-situation-reports.

[15] Kugelman, J. R., Sanchez-Lockhart, M., Andersen, K. G., et al. Evaluation of the Potential Impact of Ebola Virus Genomic Drift on the Efficacy of Sequence-Based Candidate Therapeutics. *Mbio*, 6(1), 2015.

[16] Alizon, S., Lion, S., Murall, C. L., and Abbate, J. L. Quantifying the epidemic spread of Ebola virus (EBOV) in Sierra Leone using phylodynamics. *Virulence*, 5(8):825–827, 2014.

[17] Stadler, T., Kühnert, D., Rasmussen, D. A., and du Plessis, L. Insights into the early epidemic spread of ebola in sierra leone provided by viral sequence data. *PLoS Currents*, 6, 2014.

[18] Volz, E. and Pond, S. Phylodynamic analysis of ebola virus in the 2014 sierra leone epidemic. *PLoS Currents*, 6, 2014.

[19] Matranga, C. B., Andersen, K. G., Winnicki, S., et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biology*, 15(519), 2014.

[20] Tomkins-Tinch, C., Ye, S., Metsky, H., et al. viral-ngs, 2016. URL doi:10.5281/zenodo.200428.

[21] Koster, J. and Rahmann, S. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[22] Emmett, K. J., Lee, A., Khiabanian, H., and Rabadan, R. High-resolution Genomic Surveillance of 2014 Ebolavirus Using Shared Subclonal Variants. *PLoS Currents*, 7:–, 2015.

[23] Pybus, O. G., Rambaut, A., Belshaw, R., et al. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Molecular Biology and Evolution*, 24(3):845–852, 2007.

[24] Bedford, T., Cobey, S., and Pascual, M. Strength and tempo of selection revealed in viral gene genealogies. *BMC evolutionary biology*, 11:220, 2011.

[25] Duchêne, S., Holmes, E. C., and Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc R Soc B*, 2014.

[26] Ho, S. Y. W., Phillips, M. J., Cooper, A., and Drummond, A. J. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol*, 22(7): 1561–1568, 2005.

[27] Pickett, B. E., Sadat, E. L., Zhang, Y., et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research*, 40(Database issue):D593–8, 2012.

[28] Olal, D., Kuehne, A. I., Bale, S., et al. Structure of an antibody in complex with its mucin domain linear epitope that is protective against Ebola virus. *Journal of Virology*, 86(5):2809–2816, 2012.

[29] Murin, C. D., Fusco, M. L., Bornholdt, Z. A., et al. Structures of protective antibodies reveal sites of vulnerability on Ebola virus. *Proceedings of the National Academy of Sciences*, 111(48): 17182–17187, 2014.

[30] Becquart, P., Mahlakõiv, T., Nkoghe, D., and Leroy, E. M. Identification of continuous human B-cell epitopes in the VP35, VP40, nucleoprotein and glycoprotein of Ebola virus. *PLoS ONE*, 9(6):e96360, 2014.

[31] Sanchez, A., Trappier, S. G., Stroher, U., et al. Variation in the glycoprotein and VP35 genes of Marburg virus strains. *Virology*, 240(1):138–146, 1998.

[32] Wertheim, J. O. and Worobey, M. Relaxed selection and the evolution of RNA virus mucin-like pathogenicity factors. *Journal of Virology*, 83(9):4690–4694, 2009.

[33] Gélinas, J.-F., Clerzius, G., Shaw, E., and Gatignol, A. Enhancement of replication of RNA viruses by ADAR1 via RNA editing and inhibition of RNA-activated protein kinase. *Journal of Virology*, 85(17):8460–8466, 2011.

[34] Zahn, R. C., Schelp, I., Utermöhlen, O., and von Laer, D. A-to-G hypermutation in the genome of lymphocytic choriomeningitis virus. *Journal of Virology*, 81(2):457–464, 2007.

[35] Carpenter, J. A., Keegan, L. P., Wilfert, L., O'Connell, M. A., and Jiggins, F. M. Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae). *BMC genetics*, 10:75, 2009.

[36] Folarin, O. A., Happi, A. N., and Happi, C. T. Empowering African genomics for infectious disease control. *Genome Biology*, 15(11):515, 2014.

[37] Yozwiak, N. L., Schaffner, S. F., and Sabeti, P. C. Data sharing: Make outbreak research open access. *Nature*, 518(7540):477–479, 2015.

[38] Rotmistrovsky, K. E. and Agarwala, R. BMTagger: Best Match Tagger for Removing Human Reads from Metagenomics Datasets, 2014. URL `ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/`.

[39] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[40] Yang, X., Charlebois, P., Gnerre, S., et al. De novo assembly of highly diverse viral populations. *BMC genomics*, 13:475, 2012.

[41] Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res*, 21(3):487–493, 2011.

[42] Bolger, A. M., Lohse, M., and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[43] Schmieder, R. and Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

[44] Grabherr, M. G., Haas, B. J., Yassour, M., et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[45] Charlebois, P., Yang, X., Newman, R., Henn, M., and Zody, M. V-FAT: a post-assembly pipeline for the finishing and annotation of viral genomes, 2012. URL https://www.broadinstitute.org/viral-genomics/v-fat.

[46] Lee, W.-P., Stromberg, M. P., Ward, A., et al. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*, 9(3):e90581, 2014.

[47] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[48] Novocraft Technologies. NovoAlign, 2014. URL http://www.novocraft.com/products/novoalign/.

[49] Broad Institute. Picard. URL http://broadinstitute.github.io/picard/.

[50] McKenna, A., Hanna, M., Banks, E., et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[51] DePristo, M. A., Banks, E., Poplin, R., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, 2011.

[52] Van der Auwera, G. A., Carneiro, M. O., Hartl, C., et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics*, 43:11.10.1–33, 2013.

[53] Kugelman, J. R., Wiley, M. R., Mate, S., et al. Monitoring of Ebola Virus Makona Evolution through Establishment of Advanced Genomic Capability in Liberia. *Emerging Infectious Diseases*, 21(7):1135–1143, 2015.

[54] Yang, X., Charlebois, P., Macalalad, A., Henn, M. R., and Zody, M. C. V-Phaser 2: variant inference for viral populations. *BMC genomics*, 14:674, 2013.

[55] Cingolani, P., Platts, A., Wang, L. L., et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80, 2012.

[56] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*, 29(8):1969–1973, 2012.

[57] O'Brien, J. D., Minin, V. N., and Suchard, M. A. Learning to count: robust estimates for labeled distances between molecular sequences. *Molecular Biology and Evolution*, 26(4):801–814, 2009.

[58] Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L., and Suchard, M. A. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, 28(24):3248–3256, 2012.

[59] Hasegawa, M., Kishino, H., and Yano, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[60] Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, 1994.

[61] Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Andrew Rambaut. Relaxed Phylogenetics and Dating with Confidence. *Plos Biology*, 4(5):e88, 2006.

[62] Gill, M. S., Lemey, P., Faria, N. R., et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724, 2013.

[63] Ferreira, M. A. R. and Suchard, M. A. Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics*, 36(3):355–368, 2008.

[64] Goldman, N. and Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.

[65] Lee, J. E., Fusco, M. L., Hessell, A. J., et al. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature*, 454(7201):177–182, 2008.

[66] Lee, J. E. and Saphire, E. O. Neutralizing ebolavirus: structural insights into the envelope glycoprotein and antibodies targeted against it. *Current opinion in structural biology*, 19(4): 408–417, 2009.

Chapter 3

Ebola Virus Epidemiology and Evolution in Nigeria

Preface

The Ebola virus genomes described in the previous chapter allowed us to better understand viral exchange across the Sierra Leonean border, as well as to identify patterns of selection on the virus during a period of sustained human-to-human transmission. We also identified a number of within-host variants from these data and concluded that low-frequency variants could be transmitted between hosts. As suggested in Gire et al. [1], these shared iSNVs can suggest transmission links; because low-frequency variants generally disappear or fix in a population over time [2], individuals sharing an iSNV are more likely to be close in a transmission chain, assuming the transmission bottleneck is large enough to allow transmission of multiple viruses [3]. However, as noted by Worby et al. [3], transmission dynamics and bottleneck sizes differ among viruses, and the extent to which transmission can be reconstructed using iSNVs needs to be tested on a per-virus basis. Evaluating these methods is often difficult if details about the 'true' transmission chain are

unknown.

The Ebola virus (EBOV) outbreak in Nigeria presented a unique opportunity to compare conclusions obtained from genomic data to those determined from detailed contact tracing. When EBOV entered Nigeria in 2014, detailed contact tracing was performed and ultimately helped contain the spread of the virus to just 20 individuals. In collaboration with the African Center of Excellence for Genomics of Infectious Diseases (ACEGID) at Redeemer's University Nigeria (RUN), we sequenced available patient samples and compared the resulting transmission chain to that obtained from contact tracing performed during the outbreak. The results of this investigation are published in Folarin, O. A.*, Ehichioya, D.*, Schaffner, S. F.*, Winnicki, S. M.*, Wohl, S.*, et al. *Journal of Infectious Diseases*, 2016 [4], reproduced below.

I was particularly interested in the ability of within-host variation to assist in EBOV transmission reconstruction, having recently investigated iSNVs in the Sierra Leone dataset (see Chapter 2) [5]. I performed the majority of the analysis related to the Nigeria outbreak and generated all of the figures in the resulting publication, on which I am a co-first author. Of course, this would have not been possible without the important contributions of Sarah Winnicki and Kendra West from our lab, who traveled to RUN after the outbreak to collect and sequence EBOV samples with Onikepe Folarin, Deborah Ehichioya, Philomena Eromon, and other RUN collaborators. I used detailed case reports provided by Christian Happi to reconstruct the 'true' transmission chain shown in Figure 3.2 below. Sarah Winnicki performed the difficult task of connecting samples to known cases (while maintaining patient anonymity), and Steve Schaffner calculated the likelihood of observing six mutations in 11 transmissions (see Section 3.5.8). Steve Schaffner and I were the primary

writers of the manuscript, and we are thankful to all of the individuals at RUN, the University of Lagos, the Nigeria Centre for Disease Control, and federal government of Nigeria who collected samples and patient data and, most importantly, succeeded in curtailing the spread of the outbreak.

3.1 Abstract

Containment limited the 2014 Nigerian EBOV disease outbreak to 20 reported cases and eight fatalities. We present here clinical data and contact information for at least 19 case patients, and full-length EBOV genome sequences for 12 of the 20. The detailed contact data permits nearly complete reconstruction of the transmission tree for the outbreak. The EBOV genomic data are consistent with that tree. It confirms that there was a single source for the Nigerian infections, shows that the Nigerian EBOV lineage nests within a lineage previously seen in Liberia but is genetically distinct from it, and supports the conclusion that transmission from Nigeria to elsewhere did not occur.

3.2 Introduction

The 2014 outbreak of EBOV disease (EVD) in Nigeria was one branch of the major West African epidemic that spanned 2014–2016. As of 13 March 2016, a total of 28,639 EVD cases and 11,316 deaths had been reported in 10 countries. The majority of EVD burden occurred in Liberia, Sierra Leone, and Guinea, with exported cases responsible for additional transmissions in the United States, Mali, and Nigeria, and diagnosed cases with no transmissions in the United Kingdom, Italy, Senegal, and Spain [6].

The Nigeria EVD outbreak began on 20 July 2014, when a traveler from Liberia (the index case patient), who was infected with EBOV, arrived by commercial aircraft to Murtala Muhammed International Airport in Lagos. The traveler's movement was quickly restricted, patient samples were confirmed EBOV positive by independent polymerase chain reaction (PCR) tests within days, and intensive contact tracing was conducted. The Nigeria EVD outbreak ended on 20 October 2014, when the country was declared Ebola free by the World Health Organization. During that period, 20 individuals are reported to have been infected, of whom eight died.

Despite emerging in the megacity of Lagos, the Nigeria EVD outbreak was well documented and well contained because of rapid detection of the index case and thorough contact tracing throughout the outbreak. Contact tracing provides a detailed understanding of viral spread, which is key to controlling any viral outbreak. Sequencing of patient samples can also be used to understand transmission routes and is especially important in cases where contact tracing is not available, or when contact tracing cannot completely resolve a transmission chain.

The EVD outbreak in Nigeria is unique because both genetic and contact tracing data are available. The complete transmission chain could be reconstructed with considerable confidence, and detailed clinical records were available for most patients. Viral sequencing data and sampling dates can be used to estimate general transmission patterns between patients and regions, and are used in this case to confirm and inform the transmission chain suggested by contact tracing. Comparing the two methods highlights the strengths of each, and the importance of both contact tracing and genomic sequencing during an outbreak.

We present here an account of the Nigeria 2014 EVD outbreak that includes clinical, epidemi-

ological, and viral sequence data for most of the affected patients. We also describe sequencing results generated in Nigeria and in duplicate in the United States for the purposes of both outbreak investigation and validation of viral sequencing capabilities in new laboratories.

3.3 Results

3.3.1 Clinical Data

Symptoms and outcome for all 20 patients are summarized in Tables C.1 and C.2. Their median age was 33 years (range, 26–62 years), and 55% were female. Most (65%) were <40 years of age, and most (65%) were health workers. At presentation, the most common symptoms were fever (85%), fatigue (70%), and diarrhea (65%). The pulse rate and blood pressure were within normal range in 50% of the patients, but the respiratory rate was elevated in 90% of those with available data. The common clinical syndromes documented were gastroenteritis (45%), hemorrhage (30%), and encephalopathy (15%). Of 20 patients, 12 (60%) survived, with one having a post-illness mental health complication requiring follow-up. The mean (standard deviation) duration from onset of symptoms to presentation at the ETC was three (two) days among survivors, compared with five (two) days for non-survivors. The mean duration from symptom onset to death or discharge from the Ebola treatment center (ETC) was 15 (five) days for survivors and 11 (two) days for non-survivors.

3.3.2 Sequencing Data

We prepared 16 samples from 13 of the 20 patients with confirmed EVD and discharge samples for three of them. This includes case 9, which could not be confidently matched to a sample (suspected match to E030). We prepared an additional 16 samples from suspected cases in which the sample could not be clearly associated with a particular case because of incomplete records. Because these data include retested and discharge samples, as well as incomplete information collated many months after the outbreak, we were not able to confirm that there were exactly 20 EVD cases in Nigeria. After inactivation and extraction at RUN, we divided RNA from each sample into two aliquots for independent library preparation and sequencing at RUN and the Broad Institute. Extracted RNA samples contained an average of 3.97×10^6 18S copies/mL (range, 3.28×10^4 to 2.31×10^7 copies/mL) as determined by qRT-PCR.

We prepared Nextera libraries for all 32 samples. Using the Kulesh qRT-PCR assay, we detected EBOV RNA in 18 of these samples, including two discharge samples and three samples unassociated with a particular case. After library construction, we used Kulesh qPCR to detect the presence of any EBOV copies in the libraries. Based on the results, we sequenced 23 samples using a combination of the MiSeq and HiSeq 2500 platforms (Illumina). We were able to generate assembled EBOV genomes from 12 of these samples, all from confirmed EVD cases with associated case histories. We combined the MiSeq and HiSeq sequencing data from RUN and the Broad Institute for analysis. The median sequencing coverage was 225.5× (range, 6–4,864×) (Table 3.1). Although we recorded combined sequencing data, the MiSeq data from RUN separately confirmed EBOV

reads in six of the 12 samples with assembled EBOV genomes.

3.3.3 Consensus and Within-Host Variants

We identified 17 consensus-level variants (nine synonymous, five nonsynonymous, three noncoding, all relative to the earliest EBOV sequence from the West African outbreak (GenBank accession KJ660346.2)) in EBOV genomes from the 12 sequencing-positive Nigerian samples (Table 3.2). Variants characteristic of the LB5 (Liberia sublineage 5) [7] were shared by all Nigeria EBOV genomes. The Nigerian EBOV genomes also shared three variants not common in Liberia, at positions 4,037, 17,016, and 18,754 (Table 3.2). These variants were present in all Nigerian samples sequenced, including the index case (we note that two samples did not have coverage at position 18,574). Two of these variants were unique to Nigeria, and one variant, at position 18,754, was also seen in two EBOV genomes from Liberia (accessions KT725314 and KT725261), suggesting a close relationship of the Nigeria clade to those samples. Two Nigerian samples had unique additional consensus variants.

We also identified 31 intrahost single-nucleotide variants (iSNVs) in five of the 12 EBOV genomes from Nigeria (five synonymous, five nonsynonymous, five noncoding SNPs, and 16 insertions/deletions). We sequenced each of the five samples with iSNVs at least twice from replicate libraries, and iSNV calls were concordant between libraries. Eight of these iSNVs were shared by $\geq$2 samples, and two iSNVs (positions 7,551 and 10,503), both found in sample E027, were also consensus variants in sample E030. The presence and number of iSNVs found correlated roughly with sample coverage; only samples with >100$\times$ coverage had >1 iSNV call that passed our basic filters.

Table 3.1: Sample coverage.

Sample	Case number	% Coverage	x Coverage
E001	Index	99.8%	1364
E020	2	99.5%	158
E021	3	99.8%	520
E023	4	99.7%	525
E024	5	99.8%	4864
E027	6	99.5%	159
E029	7	99.7%	474
E033	8	82.4%	6
E030	9	99.8%	292
E039	10	90.4%	8
E076	11	99.1%	25
E130	13	99.1%	14

% Coverage = percentage of bases with $\geq 1\times$ coverage
$\times$ Coverage = median depth of coverage

Table 3.2: Consensus SNPs seen in Nigeria[a].

Position	Ref Allele	Alt Allele	Type	Gene	Substitution	Lineage	Count
800	C	T	Missense	NP	R111C	SL2	12
1849	T	C	Silent	NP	D460D	SL1	12
2895	C	T	Non-coding	–	–		1 (E020)
3336	A	G	Missense	VP35	N70D		1 (E020)
3920	G	A	Silent	VP35	Q264Q		1 (E020)
4037[b]	T	C	Silent	VP35	I303I		12
6056	A	C	Silent	GP	I6I	LB5	12
6283	C	T	Missense	GP	A82V	SL1	11[c]
7551	T	C	Missense	GP	V505A		1 (E030)
8928	A	C	Silent	VP30	P140P	SL2	12
10503	A	G	Silent	VP24	G53G		1 (E030)
11201	A	G	Non-coding	–	–		1 (E020)
15963	G	A	Silent	L	K1461K	SL2	12
16514	G	A	Missense	L	S1645N	LB5	12
17016[b]	C	T	Silent	L	S1812S		12
17142	T	C	Silent	L	F1854F	SL2	12
18754[b]	A	T	Non-coding	–	–		10[d]

[a] All variants and positions are relative to the KJ660346.2 Guinea genome from early in the outbreak. The lineage column includes previously published clade-defining SNPs ancestral to the Nigeria lineage.

[b] These three SNPs are novel to Nigeria (except 18,754, which is shared by two Ebola virus genomes from Liberia) and are shared by all Nigerian samples.

[c] No coverage in sample E033.

[d] No coverage in sample E033 or E039.

3.3.4 Phylogenetic Tree

To better understand the evolutionary relationship between the EVD outbreak in Nigeria and the West African outbreak as a whole, we created a maximum likelihood tree (Figure 3.1). The tree confirms that the EVD outbreak in Nigeria was due to a single introduction from Liberia, as suggested by contact tracing. More specifically, the EBOV genomes from Nigeria are descendants of the LB5 clade in Liberia [7]. No EBOV sequences yet sampled outside Nigeria descend from the Nigerian EBOV isolates [1, 8–12], indicating containment of EVD cases in Nigeria within the larger outbreak, as also suggested by contact tracing.

3.3.5 Reconstructed Transmission Tree

Given the phylogenetic tree of the sampled viruses, along with their dates, it is possible to infer at least the outlines of the chain of transmission from one patient to another (Figure 3.2A). Ten Nigerian EBOV have identical consensus sequences, suggesting that these sequences are closely connected by direct transmissions. Date information identifies sample E001, the index case, as the earliest-sampled case in Nigeria (collection date: 22 July 2014). Of the other nine identical genomes, seven have collection dates from 4 August 2014–8 August 2014.

The close proximity of the sample collection dates to each other suggests that each of the corresponding case patients was infected by the index case patient (i.e., it is unlikely that an individual presenting symptoms on 8 August 2014 would have been infected <4 days previously) [13]. The remaining two cases with viral genomes identical to the index case are dated 15 August and 1

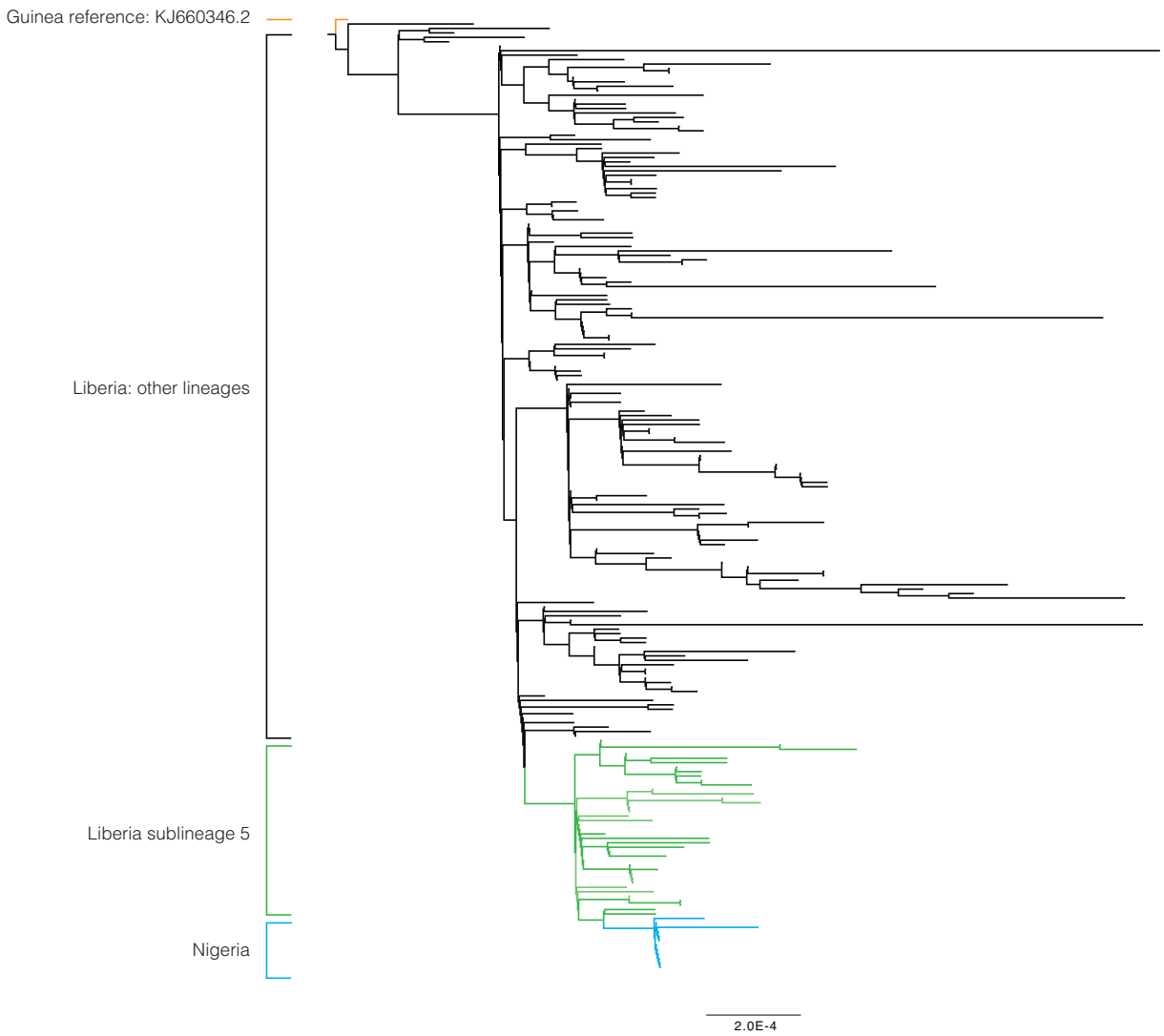


Figure 3.1: Maximum likelihood tree. Phylogenetic analysis confirms a single introduction of Ebola virus into Nigeria from Liberia and places all Nigerian sequences as descendents of Liberia sublineage 5. Two Liberia sublineage genomes (GenBank accessions: KT725314 and KT725261) cluster closely with Nigerian samples owing to a shared variant at position 18,754. Scale bar indicates nucleotide substitutions per site.

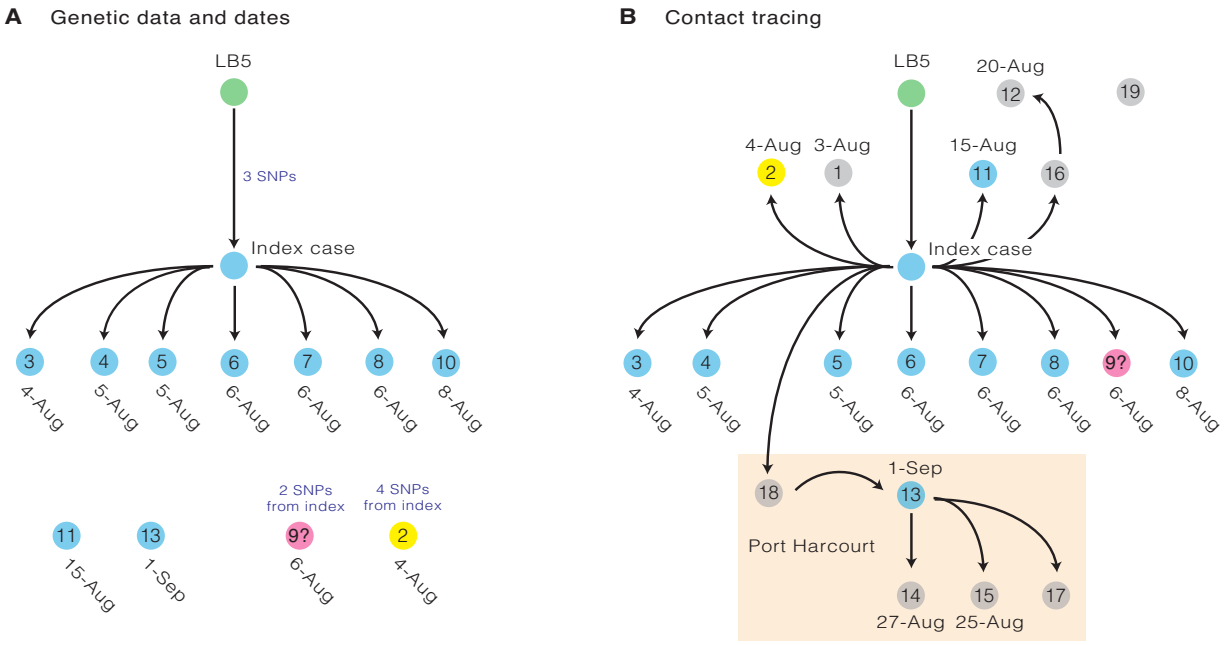


Figure 3.2: Transmission tree. **(A)** Transmission reconstructed from of Ebola virus genome sequence and sample dates only. Arrows indicate likely transmission; cases not connected to arrows cannot be placed within the transmission tree given the available data. LB5, Liberia sublineage 5 reference. **(B)** Transmission reconstructed from contact tracing only. Contact tracing provides more precise information, but is not always available. Samples were collected in Lagos, Nigeria, unless otherwise identified. Each case is labeled with its sample collection date; cases not connected to sequenced samples are labeled with date of hospitalization. Samples are colored by consensus sequence (i.e., samples with identical viral genomes are similarly colored). Cases in gray are those for which genetic data are not available.

September, and these patients therefore may have been infected by one of the earlier case patients. The presence of additional SNPs in the viral genomes corresponding to cases 2 and 9 make it difficult to place these samples within the transmission chain. However, case 6 has an iSNV at each of the two case 9 SNP positions (position 7,551, 21% minor allele frequency; position 10,503, 16% minor allele frequency), suggesting that these two cases are closely linked.

In the limited Nigeria EVD outbreak, it was also possible to reconstruct a nearly complete transmission chain based on contact tracing alone (Figure 3.2B). Such a reconstruction is feasible in this case because (1) EBOV spreads primarily through direct contact, (2) there were few cases (multiple exposures were uncommon), and (3) intensive efforts were made to trace and monitor all suspected contacts. The contact tracing information resulted in a transmission tree similar to that suggested by genetic data, with the index case responsible for a majority of transmissions. This data also revealed that one individual (case 18) traveled from Lagos to Port Harcourt while infected with EBOV, where he acted as the index patient in a small secondary outbreak containing four additional EVD cases.

3.4 Discussion

The 2014 Nigeria outbreak is unusual for an EVD outbreak in the detailed information available about its development: we have both a good reconstruction of the transmission chain of 20 patients, and viral genomic data from most cases in the chain. The completeness of the record reflects the public health situation: Nigeria was prepared for the arrival of EBOV and was able to im-

plement thorough contact tracing promptly after the index case was diagnosed, while the number of cases was still small. That effort was critical in containing the outbreak, but it is also very helpful in reconstructing its details afterward. Combined with sequence data, the transmission chain helps us interpret the changes occurring in the virus, because it generally lets us pinpoint where in the chain each new mutation actually occurred.

Viewed by itself, sequence data can serve to provide a broad picture of an outbreak, and that is true of this EVD outbreak. This capability is obviously useful when contact tracing is absent or incomplete, as is usually the case with epidemics. In the 2014 Nigeria outbreak, sequencing alone makes it clear that the entire outbreak stemmed from a single introduction of EBOV into the country. It also places the Nigerian outbreak in its larger context, identifying a particular branch of the Liberian LB5 lineage of EBOV as the source and showing that the Nigerian lineage did not spread into other countries.

Identifying individual links in the transmission chain is usually beyond the resolution of sequence data, however, and requires contact tracing in the field. The resolution of genomic data is limited because new variants arise less often than new cases, meaning that many cases will be genetically indistinguishable. This can be seen in our data in Figure 3.2A, in which multiple successive links in the chain share identical genomes. In addition, when mutations do occur, >1 can arise in a single patient, making genetic distance an imperfect guide to the number of transmission links that have occurred. Thus, most of the cases infected directly by the index patient in Nigeria had identical genomes, but one case (case 4) differed by four mutations, even though it too resulted from a single transmission. Contact tracing (Figure 3.2B) — when it is available — does not suffer

from such limitations.

Within-host variants (iSNVs) that are shared between patients can provide a more detailed picture of transmission routes, but our data point out some important caveats about their usefulness. First, detection of iSNVs requires deep sequencing of good-quality samples, and that is not always possible: deep enough sequencing could be achieved for only two-thirds of our sequenced samples. Second, even when iSNV data are available, it may not all be meaningful. Some of the iSNVs we observed have previously been documented in unrelated data sets from Sierra Leone and Liberia [1, 5, 7]; these included all eight of the shared iSNVs. Most of our iSNVs, including most shared iSNVs, were low-frequency frameshift insertions or deletions. Because they can disrupt protein structure, they are unlikely to be transmitted. More likely, these iSNVs represent either recurring mutations in highly mutable regions of the EBOV genome or sequencing errors, especially because many of them occur in homopolymer regions. In either case, their value for determining transmission chains is uncertain. More research is necessary to fully make use of within-host genomic data in understanding transmission, including better sequencing coverage for all samples and improved methods to identify false-positives.

One aspect of our genomic data that is slightly surprising is the distribution of new variants, which is not at all uniform. Our sequenced samples include the results of 11 transmissions from the index case. Nine of these produced no new consensus SNPs, one produced four new SNPs, and one produced two (Figure 3.2A). This clustering of mutations in certain samples suggests the possibility that the mutation rate was not uniform across all of the cases. This is no more than a possibility, though, because the clustering is not statistically significant (p=0.07).

Also puzzling is a pair of variants that were seen twice, once as consensus SNPs (in case 9) and once as iSNVs (in case 6). Based on sample dates and contact data, both of these patients were infected by the index patient, so presumably they inherited these variants from that patient. We do not, however, find them in the sample from the index patients, either as consensus SNPs or as iSNVs, despite high sequencing depth. Nor do they appear as consensus SNPs in the other cases derived from the index case, or as iSNVs in the one other case that was deeply sequenced and was sampled around the same time as samples 6 and 9. The explanation may simply be that the variants were present in the index patient but at too low a frequency for us to detect. It is also possible that their frequency changed in the index patient between the time he was sampled and transmission to the other cases, or that they differed across tissues within the patient. Better understanding of the dynamics of within-host evolution and transmission, and of our power to detect iSNVs, would help clarify this issue.

The genomic data were invaluable in revealing what was happening to the virus during the outbreak, but it would have been even more informative had samples been of uniformly high quality. Many samples did not produce whole-genome assemblies because of poor sample quality, and a third of those that did could not be used to detect iSNVs. This highlights the importance of rapid sequencing in clinical settings during outbreaks, with well-established sample collection and processing protocols. Although at the time of the outbreak sequencing was not yet ready on site, sequencing capability is now becoming increasingly available throughout many regions. With high-throughput deep sequencing now being routinely performed by ACEGID at RUN, high-resolution pathogen information can now be generated to elucidate outbreak dynamics and response, both in

Nigeria and throughout West Africa.

Data handling could similarly benefit from good protocols established in advance. In the case of the data presented here, clinical and contact data were separated from sequence data, and the correspondence between them had to be established post hoc, a process that was both laborious and uncertain. In an outbreak setting, keeping track of different kinds of data is not the highest priority, but valuable information can be lost as a result. Having a system for collecting and maintaining both clinical and laboratory data established in advance would be very helpful.

3.5 Methods

3.5.1 Management of Contacts and Cases of EVD

The index case patient presented to a private hospital in Lagos on 20 July 2014 with fever and body weakness, denied contact with known EVD cases or funeral attendance, and was treated with antimalarial drugs and analgesics. Over the next three days, the patient's condition worsened (fever escalated, and vomiting and diarrhea persisted), and EVD was suspected. Filovirus PCR testing was conducted at Lagos University Teaching Hospital, and on 23 July the index case was reported as filovirus positive. Samples were then shared with Redeemer's University (RUN) for EBOV-specific PCR testing, which was confirmed on 25 July 2014. The index patient died on 25 July 2014.

All persons who were exposed to the index patient and their contacts were traced, placed under surveillance, and monitored for clinical features of EVD. If contacts exhibited fever or other symptoms, they were admitted into the Ebola treatment center (ETC) as suspected case patients;

blood samples were then collected and tested with reverse-transcription (RT) PCR for presence of EBOV at both Lagos University Teaching Hospital and RUN. Patients who tested positive with RT-PCR were moved to the confirmed ward of the ETC. This combination — history of contact with an EVD case patient, presentation with symptoms, and RT-PCR evidence of EBOV infection — defined a confirmed case. Each patient was counseled on the need for $\geq$4 L of oral rehydration solution daily. Treatment was started with antibiotics because of their immunosuppression and antimalarials because of the endemicity of malaria in Nigeria. Patients were also placed on a regimen of nutritional supplements and vitamins. The only analgesic administered was paracetamol. Injectables and invasive procedures were avoided unless patients were too ill or weak to take oral rehydration solution.

Infection prevention and control procedures and protocols were strictly adhered to in patient management. Before discharge, patients were confirmed negative for EVD by RT-PCR. When discharged, they were decontaminated before being allowed to leave the ETC and were not allowed to take clothing or other personal items. Replacement clothes, footwear, and basic personal effects were provided by family or the ETC, depending on individual circumstances.

3.5.2 Data Collection and Review

ETC case management, clinical data, and laboratory data of all confirmed EVD cases identified between 20 July and 30 September were reviewed by qualified medical professionals in the case management team. The following case data were compiled: sociodemographic (age, sex, occupation, and city of residence), clinical (respiratory rate, pulse rate, blood pressure, presenting symptoms,

signs, syndromes, and outcome), laboratory (RT-PCR), and administrative data (date of symptoms onset, duration of symptoms, and length of stay).

Each patient's exposure history, presenting symptoms, history of presenting symptoms, course of illness, excerpts of clinical management, and illness outcome were abstracted from medical records or contact tracing interview notes (including suspect evacuation forms, case investigation forms, laboratory request and report forms, clinical notes and charts, and contact tracing interview notes) and summarized as case histories.

3.5.3 Sample Collection and Processing

Samples from patients with suspected EVD were shipped both to the virology laboratory at Lagos University Teaching Hospital for diagnostics and to ACEGID at RUN for diagnostics and sequencing. Whole-blood samples shipped to RUN were inactivated with AVL buffer (Qiagen) or TRIzol LS reagent (Life Technologies) in a 4:1 ratio, both according to the manufacturer's protocol. Inactivated samples were stored in a -20°C freezer. AVL buffer and TRIzol LS reagent have been used extensively in virus inactivation including for EBOV [1, 14–18]. Samples inactivated in AVL buffer were extracted using the QIAamp Viral RNA Mini Kit extraction protocol (Qiagen), according to the manufacturer's protocol. Samples inactivated in TRIzol reagent were extracted using chloroform modified with an AVL buffer inactivation and QIAamp Viral RNA Mini Kit extraction protocol. Following this modified protocol, 140 μL of chloroform was added to 1 mL of a TRIzol-inactivated sample. After vortex and centrifugation, 200 μL of the aqueous phase was transferred to a tube with 700 μL of AVL buffer without carrier RNA added. The sample was then processed according to

the manufacturer's protocol for extraction, using the QIAamp Viral RNA Mini Kit. Extracted RNA samples were divided into aliquots for sequencing at both RUN and the Broad Institute of MIT and Harvard. Samples destined for the Broad Institute were shipped on dry ice and subsequently stored at −80°C.

3.5.4 Diagnostics Performed at RUN

EBOV-specific diagnostic tests were performed on the suspected EBOV samples at RUN with RT-PCR using the SuperScript III One-Step RT-PCR System with Platinum Taq High Fidelity DNA Polymerase (Life Technologies). The 25-μL assay mix included 5 μL of RNA, KGH primer set [1] at a 250 nmol/L final concentration (forward, GTC GTT CCA ACA ATC GAG CG; reverse, CGT CCC GTA GCT TTR GCC AT), 12.5 μL of 2× Reaction Mix and 0.5 μL of SuperScript III RT/Platinum Taq High Fidelity Enzyme Mix. The cycling conditions were 60°C for 20 minutes and 94°C for five minutes, followed by 35 cycles of 94°C, 58°C, and 68°C for 15 seconds each, with a final extension at 68°C for two minutes. RT-PCR was performed on an Eppendorf Mastercycler thermocycler. The samples were analyzed on 1.5% agarose gel, and visual results were recorded.

3.5.5 Quantitative RT-PCR Performed at RUN and the Broad Institute

To assess sample quality, extracted RNA was quantified using quantitative RT-PCR (qRT-PCR) for both EBOV and human ribosomal RNA (18S). RNA selected for sequencing was quantified using the Power SYBR Green RNA-to-Ct 1-Step qRT-PCR assay (Life Technologies). The Kulesh assay protocol was adapted from a probe-based quantitative PCR (qPCR) assay to a SYBR qPCR assay by

omitting the probe [19]. The 10-μL assay mix included 3 μL of RNA, 0.3 μmol/L primer Kulesh forward (TCT GAC ATG GAT TAC CAC AAG ATC), 0.3 μmol/L Kulesh reverse (GGA TGA CTC TTT GCC GAA CAA TC), 5 μL of 2$\times$ Power SYBR Green RT-PCR Mix and 0.08 μL of RT Enzyme Mix (Life Technologies). The cycling conditions were 48°C for 30 minutes and 95°C for 10 minutes, followed by 45 cycles of 95°C for 15 seconds and 60°C for 30 seconds with a melt curve of 95°C for 15 seconds, 55°C for 15 seconds, and 95°C for 15 seconds. qRT-PCR was performed on the LightCycler 96 (Roche) instrument at both RUN and the Broad Institute. Synthetic oligonucleotide amplicons were prepared as a standard to quantify the viral copy number in the qRT-PCR assays. These amplicons represent a portion of the EBOV segment within the L gene as a template for PCR. The amplicons were cleaned using AMPure XP beads (Beckman Coulter Genomics) and quantified by the TapeStation system (Ambion). Amplicon concentrations were converted to EBOV copies per microliter for quantification.

3.5.6 RNA Processing and Library Preparation

DNA was depleted from the RNA samples using TURBO DNase (Ambion), and host ribosomal RNA was then depleted from the samples using an RNase H selective depletion method described elsewhere [1, 20, 21]. Complementary DNA was then synthesized from the resulting depleted RNA, Nextera XT libraries were constructed, and Illumina sequencing was carried out according to methods described elsewhere [1, 22], with the modification that Nextera libraries were generated using 16–18 cycles of PCR. Samples were sequenced on the MiSeq platform at RUN, and on both the MiSeq and HiSeq 2500 platforms (Illumina) at the Broad Institute.

3.5.7 EBOV Genome Assembly and Analysis

Raw sequencing reads from all sequencing runs were processed together and assembled using the viral-ngs pipeline (version 1.0.0) [5, 23] with mostly default parameters. Reads from two flow cells were not included owing to suspected contamination. Two parameters were varied from defaults: 'assembly_min_length_fraction_of_reference' was set to 0.8, and 'assembly_min_unambig' was set to 0.7. Sequence assemblies are available from GenBank and reads available from the sequence read archive, accessible under BioProject PRJNA316870.

Consensus variants were called using a custom pipeline and annotated using the program SnpEff (version 4.1) [24]. Multiple alignments were performed using MAFFT software (version 7.017) [25, 26] with default parameters. Within-host variants were identified as part of the viral-ngs pipeline with default minimum read and strand bias filters.

The maximum likelihood tree was produced using IQ-TREE software (version 1.3.13) [27], a TIM+I (a transitional model with a proportion of invariable sites) substitution model selected by ModelFinder (implemented in IQ-TREE), and 1000 bootstrap replicates. Liberian EBOV sequences included all genomes publicly available on GenBank as of 17 February 2016.

3.5.8 Data Analysis

As noted in the Discussion, new single-nucleotide polymorphisms (SNPs) were observed to be clustered, with six SNPs appearing in one sample, two in another, and none in the remaining nine samples. To determine whether this was unlikely given a uniform mutation rate per transmission, a p

value was calculated as follows. From the transmission tree, the sequenced cases represent a minimum of 11 transmissions from the index case. Assume that new SNPs in a transmission occur in a Poisson process at an unknown rate, μ_s. For a given μ_s, we calculate the probability of seeing four new SNPs in $\geq$1 case and then integrate over all values of μ_s, weighting by the probability of observing six SNPs in 11 transmissions. That is,

$$p = \frac{\int p(S_t = 6 \mid \mu_s)(1 - p(S_s < 4|\mu_s)^{N_t})d\mu_s}{\int p(S_t = 6 \mid \mu_s)d\mu_s}$$

where S_t is the total number of new SNPs, S_s is the number of new SNPs seen in a single case, and N_t is the number of transmissions. The first probability is the Poisson probability density function, $p(S_t \mid \mu_s) = ((\mu_s N_t)^{S_t} e^{-\mu_s N_t})/S_t!$, and the second is the cumulative distribution function, $e^{-\mu_s} \sum_{i=0}^{N_t - 1} (\mu_s^i / i!)$.

3.6 Acknowledgements

We would like to thank Mike Lin and Yifei Men at DNAnexus for their engineering work to assist with analysis of data generated by ACEGID at RUN. We also thank the RUN management for the support provided to ACEGID staff during the 2014 EVD outbreak in Nigeria, members of the Emergency Operations Center in Nigeria during the outbreak, and the federal government of Nigeria.

This work was supported by grants from Illumina Corporation and the United States Agency for International Development (grant OAA-G-15-00001), the National Institutes of Health (NIH) (grant 5U01HG007480-03), the World Bank (ACE 019), the Bill and Melinda Gates Foundation

(grant OPP1123407), and the Howard Hughes Medical Institute (to P.C.S.). Sequencing and analysis work at the Broad Institute was supported by federal funds from the National Institute of Allergy and Infectious Diseases, NIH, Department of Health and Human Services (under grant U19AI110818).

3.7 References

[1] Gire, S. K., Goba, A., Andersen, K. G., et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[2] Stack, J. C., Murcia, P. R., Grenfell, B. T., Wood, J. L. N., and Holmes, E. C. Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proceedings of the Royal Society B: Biological Sciences*, 280(1750), 2013.

[3] Worby, C. J., Lipsitch, M., and Hanage, W. P. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. *American journal of epidemiology*, 186(10):1209–1216, 2017.

[4] Folarin, O. A., Ehichioya, D., Schaffner, S. F., et al. Ebola Virus Epidemiology and Evolution in Nigeria. *The Journal of Infectious Diseases*, 214(suppl 3):S102–S109, 2016.

[5] Park, D. J., Dudas, G., Wohl, S., et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526, 2015.

[6] Organization, W. H. Ebola Situation Reports - 16 March 2016, 2016. URL http://apps.who.int/ebola/current-situation/ebola-situation-report-16-march-2016.

[7] Ladner, J. T., Wiley, M. R., Mate, S., et al. Evolution and Spread of Ebola Virus in Liberia, 2014-2015. *Cell Host and Microbe*, 18(6):659–669, 2015.

[8] Baize, S., Pannetier, D., Oestereich, L., et al. Emergence of Zaire Ebola Virus Disease in Guinea — Preliminary Report. *New England Journal of Medicine*, 371(15):1418–1425, 2014.

[9] Tong, Y.-G., Shi, W.-F., Liu, D., et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*, 524(7563):93–96, 2015.

[10] Carroll, M. W., Matthews, D. A., Hiscox, J. A., et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature*, 524(7563):97–U201, 2015.

[11] Simon-Loriere, E., Faye, O., Faye, O., et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*, 524(7563):102–U210, 2015.

[12] Kugelman, J. R., Wiley, M. R., Mate, S., et al. Monitoring of Ebola Virus Makona Evolution through Establishment of Advanced Genomic Capability in Liberia. *Emerging Infectious Diseases*, 21(7):1135–1143, 2015.

[13] Chowell, G. and Nishiura, H. Transmission dynamics and control of Ebola virus disease (EVD): a review. *BMC medicine*, 12:196, 2014.

[14] Günther, S., Asper, M., Röser, C., et al. Application of real-time PCR for testing antiviral compounds against Lassa virus, SARS coronavirus and Ebola virus in vitro. *Antiviral research*, 63 (3):209–215, 2004.

[15] Grard, G., Biek, R., Tamfum, J.-J. M., et al. Emergence of divergent Zaire ebola virus strains in Democratic Republic of the Congo in 2007 and 2008. *The Journal of Infectious Diseases*, 204 Suppl 3:S776–84, 2011.

[16] Kobinger, G. P., Leung, A., Neufeld, J., et al. Replication, pathogenicity, shedding, and transmission of Zaire ebolavirus in pigs. *The Journal of Infectious Diseases*, 204(2):200–208, 2011.

[17] Hoenen, T., Jung, S., Herwig, A., Groseth, A., and Becker, S. Both matrix proteins of Ebola virus contribute to the regulation of viral genome replication and transcription. *Virology*, 403 (1):56–66, 2010.

[18] Blow, J. A., Mores, C. N., Dyer, J., and Dohm, D. J. Viral nucleic acid stabilization by RNA extraction reagent. *Journal of virological methods*, 150(1-2):41–44, 2008.

[19] Trombley, A. R., Wachter, L., Garrison, J., et al. Comprehensive panel of real-time TaqMan polymerase chain reaction assays for detection and absolute quantification of filoviruses, arenaviruses, and New World hantaviruses. *The American journal of tropical medicine and hygiene*, 82(5):954–960, 2010.

[20] Matranga, C. B., Andersen, K. G., Winnicki, S., et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biology*, 15(519), 2014.

[21] Morlan, J. D., Qu, K., and Sinicropy, D. V. Selective Depletion of rRNA Enables Whole Transcriptome Profiling of Archival Fixed Tissue. *PLoS One*, 7:e42882, 2012.

[22] Adiconis, X., Borges-Rivera, D., Satija, R., et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature methods*, 10(7):623–629, 2013.

[23] Tomkins-Tinch, C., Ye, S., Metsky, H., et al. viral-ngs, 2016. URL doi:10.5281/zenodo.200428.

[24] Cingolani, P., Platts, A., Wang, L. L., et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80, 2012.

[25] Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[26] Katoh, K. and Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, 2013.

[27] Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015.

Chapter 4

Zika Virus Epidemiology and Evolution in the Americas

Preface

Zika virus (ZIKV) was first reported in the Americas (specifically, in Brazil) in 2015, and soon spread throughout the region [1–4]. Due to the apparent connection between the virus and birth defects [5–7], the Zika epidemic soon became the focus of numerous sequencing studies and public health initiatives [1, 8–11]. Our lab's work on ZIKV is published in Metsky, H.C.*, Matranga, C.B.*, Wohl, S.*, Schaffner, S.F.*, et al. *Nature*, 2017 [11]. I am a co-first author on this paper and, during the course of this project, I worked with Hayden Metsky to analyze and troubleshoot sequencing results. I also investigated within-host variation, prepared some samples for sequencing (with Chris Matranga), and was involved in conversations about how to draw conclusions from limited data. And through my work helping to prepare figures for publication, I became intimately involved in the discussion for what each should show.

In the process of analyzing ZIKV and responding to the outbreak, we performed an in-depth

exploration of sequencing methods and iSNV identification; these are analyses I was primarily responsible for, and they are summarized in Figure 4.1A, Figure 4.4, and Table D.3 (note that all of this analysis was done in a close partnership with Hayden Metsky). The chapter below is adapted from our publication, with special emphasis on the sections in which I was most involved (see Sections 4.3.1 and 4.3.4).

4.1 Abstract

Although the recent ZIKV epidemic in the Americas and its link to birth defects have attracted a great deal of attention [5, 6], much remains unknown about ZIKV disease epidemiology and ZIKV evolution, in part owing to a lack of genomic data. Here we address this gap in knowledge by using multiple sequencing approaches to generate 110 ZIKV genomes from clinical and mosquito samples from 10 countries and territories, greatly expanding the observed viral genetic diversity from this outbreak. We analyzed the timing and patterns of introductions into distinct geographic regions; our phylogenetic evidence suggests rapid expansion of the outbreak in Brazil and multiple introductions of outbreak strains into Puerto Rico, Honduras, Colombia, other Caribbean islands, and the continental United States. We find that ZIKV circulated undetected in multiple regions for many months before the first locally transmitted cases were confirmed, highlighting the importance of surveillance of viral infections. We identify mutations with possible functional implications for ZIKV biology and pathogenesis, as well as those that might be relevant to the effectiveness of diagnostic tests.

4.2 Introduction

Since its introduction into the Americas, mosquito-borne ZIKV (family: Flaviviridae) has spread rapidly, causing hundreds of thousands of cases of ZIKV disease, as well as ZIKV congenital syndrome and probably other neurological complications [5, 6, 12]. Phylogenetic analysis of ZIKV can reveal the trajectory of the outbreak and detect mutations that may be associated with new disease phenotypes or affect molecular diagnostics. Despite the 70 years since its discovery and the scale of the recent outbreak, however, fewer than 100 ZIKV genomes have been sequenced directly from clinical samples. This is due in part the technical challenge of sequencing ZIKV, which is difficult because the viral load is relatively low compared to viruses with acute infections like Ebola virus (EBOV) [13], or even compared to other flaviviruses such as Dengue [14].

Specifically, while EBOV titers often range from 10^5–10^7 cp/ml [13], ZIKV titer typically ranges from 10^2–10^5 cp/ml [15, 16] (possibly up to 10^5 cp/ml in urine [16]). Additionally, ZIKV titer in most bodily fluids drops rapidly during the first week of symptomatic infection [16, 17], making it difficult to detect and sequence unless a patient is diagnosed shortly after he or she becomes symptomatic. Since ZIKV often causes short-lived, non-specific, flu-like symptoms [18], it is not always immediately suspected and diagnosed. The lack of available ZIKV sequences my also be explained by loss of RNA integrity in samples collected and stored without sequencing in mind. Culturing the virus increases the material available for sequencing but can result in genetic variation that is not representative of the original clinical sample.

4.3 Results

4.3.1 Two Approaches for Zika Virus Sequencing

We sought to gain a deeper understanding of the viral populations underpinning the ZIKV epidemic by extensive genome sequencing of the virus directly from samples collected as part of ongoing surveillance. We initially pursued unbiased metagenomic sequencing to capture both ZIKV and other viruses known to be co-circulating with ZIKV [19]. In most of the 38 samples examined by this approach there proved to be insufficient ZIKV RNA for genome assembly, but it still proved valuable to verify results from other methods. Metagenomic data also revealed sequences from other viruses, including 41 likely novel viral sequence fragments in mosquito pools (Table D.1). In one patient we detected no ZIKV sequence but did assemble a complete genome from dengue virus (type 1), one of the viruses that co-circulates with and presents similarly to ZIKV [20].

To capture sufficient ZIKV content for genome assembly, we turned to two targeted approaches for enrichment before sequencing: multiplex PCR amplification [21] and hybrid capture [22]. We sequenced and assembled complete or partial genomes from 110 samples from across the epidemic, out of 229 attempted (221 clinical samples from confirmed and possible ZIKV disease cases and eight mosquito pools; Table 4.1). This dataset, which we used for further analysis, includes 110 genomes produced using multiplex PCR amplification (amplicon sequencing) and a subset of 37 genomes produced using hybrid capture (out of 66 attempted). Because these approaches amplify any contaminant ZIKV content, we relied heavily on negative controls to detect artefac-

Table 4.1: Samples and genomes by region.

Country or territory	Samples	Samples with metagenomic data	Amplicon sequencing genomes	Hybrid capture genomes	Total genomes
Brazil	53	12	27	7	**27**
Colombia	20	0	4	2	**4**
Dominican Republic	45	7	30	9	**30**
Guatemala/El Salvador	3	0	1	0	**1**
Haiti	4	0	1	0	**1**
Honduras	20	6	18	8	**18**
Jamaica	20	0	5	0	**5**
Martinique	3	0	1	0	**1**
Puerto Rico	15	0	3	1	**3**
Continental US	36	12	20	10	**20**
Other	10	1	0	0	**0**
Total	229	38	110	37	**110**

Sample source information and sequencing results for 229 clinical and mosquito pool samples. Continental United States includes eight mosquito pool samples; all others are clinical samples from the Americas. In the final column, genomes generated by both methods are counted only once. 'Other' includes regions without a ZIKV genome included in downstream analysis.

tual sequence, and we established stringent, method-specific thresholds on coverage and completeness for calling high-confidence ZIKV assemblies (Figure 4.1A). We typically include a water sample (or a sample from a completely different virus, such as EBOV) when preparing samples for sequencing, and thus used ZIKV content in these negative controls as a proxy for the level of contamination present in that particular batch of samples. We used this principle to define inclusion thresholds for ZIKV assemblies generated in this study. Given the significant differences between the hybrid capture and amplicon sequencing methods, we developed a different set of criteria for each method.

For the hybrid capture method, we identified the highest number of unambiguous (non-'N') bases observed in any negative control across all sequencing runs, $B_{NC} \approx 2{,}000$, and set the thresh-

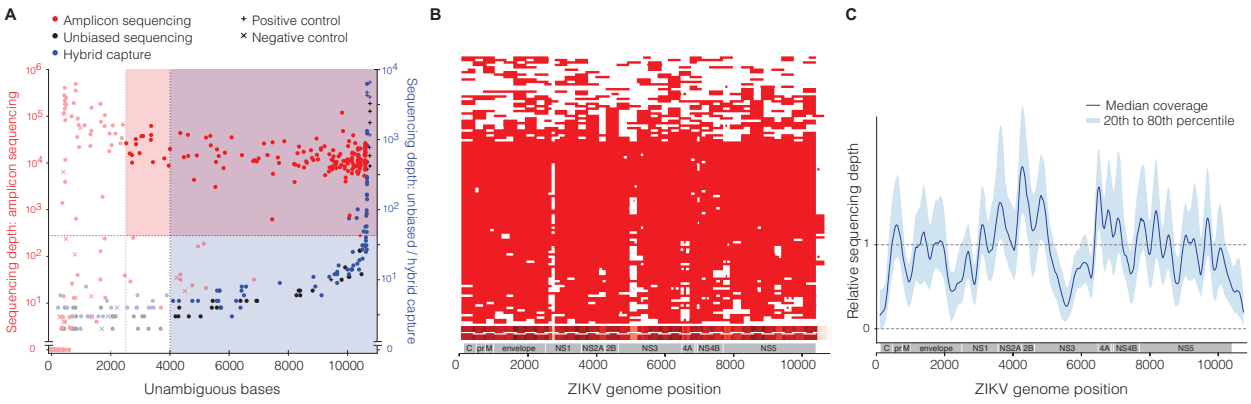


Figure 4.1: Sequence data from clinical and mosquito samples. (**A**) Thresholds used to select samples for downstream analysis. Each point is a replicate. Red and blue shading: regions of accepted amplicon sequencing and hybrid capture genome assemblies, respectively. Not shown: hybrid capture positive controls with depth >10,000×. (**B**) Amplicon sequencing coverage by sample (row) across the ZIKV genome. Red, sequencing depth ≥100×; heatmap (bottom) sums coverage across all samples. White horizontal lines on heatmap, amplicon locations. (**C**) Relative sequencing depth across hybrid capture genomes.

old at approximately $2 * B_{NC} = 4{,}000$. We could have set a unique threshold for every run, but opted for this more conservative measure. For the amplicon sequencing method, we noticed that sequencing assemblies were roughly binary in their median sequencing depth: the best samples had a depth well over 1,000×, and poor samples generally had a coverage under 100×. To set the exact threshold of 275×, we used data from positive controls and ensured that the assembly of every positive control would pass our threshold. As an additional precaution, we required 2,500 unambiguous bases (roughly 25% of the genome) in sequences produced via amplicon sequencing (see also Methods, Section 4.5).

While using negative controls to set inclusion thresholds for assemblies may seem a simple and practical task, it is important to note that we had not previously employed such thresholds for sequencing EBOV or other viruses. This is largely because contamination is much less of an issue

in the absence of amplification prior to sequencing, or when dealing with samples with higher viral content. That said, we have since used inclusion thresholds (the method, not the specific numerical values) when sequencing other viruses, such as mumps, as described in the following chapter.

After applying these thresholds to select genomes for downstream analysis, we calculated completeness and coverage, shown in Figure 4.1B–C; the median fraction of the genome with unambiguous base calls was 93%. Per-base discordance between genomes produced by the two methods was 0.017% across the genome, 0.15% at polymorphic positions, and 2.2% for minor allele base calls. Patient sample type (urine, serum, or plasma) made no significant difference to sequencing success in our study (Figure D.1).

4.3.2 Presence of Zika Virus in the Americas Prior to Clinical Diagnosis

To investigate the spread of ZIKV in the Americas, we performed a phylogenetic analysis of the 110 genomes from our dataset together with 64 published genomes available on NCBI GenBank and in Faria et al. [9] and Grubaugh et al. [10] (Figure 4.2A). Our reconstructed phylogeny (Figure 4.2B), which is based on a molecular clock (Figure D.2), is consistent with the outbreak having originated in Brazil [1]: Brazil ZIKV genomes appear on all deep branches of the tree, and their most recent common ancestor is the root of the entire tree. We estimate the date of that common ancestor to have been in early 2014 (95% credible interval (CI): August 2013–July 2014). The shape of the tree near the root remains uncertain (that is, the nodes have low posterior probabilities) because there are too few mutations to distinguish the branches. This pattern suggests rapid early spread of the outbreak, consistent with the introduction of a new virus to an immunologically naive population.

ZIKV genomes from Colombia (n=10), Honduras (n=18), and Puerto Rico (n=3) cluster within distinct, well-supported clades. We also observed a clade consisting entirely of genomes from patients infected in one of three Caribbean countries (the Dominican Republic, Jamaica, and Haiti) or the continental United States, containing 30 of 32 genomes from the Dominican Republic and 19 of 20 from the continental United States. We estimated the within-outbreak substitution rate to be 1.15×10^{-3} substitutions per site per year (95% CI: 9.78×10^{-4} to 1.33×10^{-3}), similar to prior estimates for this outbreak [1]. This is 1.3–5 times higher than reported rates for other flaviviruses [23], but is measured over a short sampling period, and therefore may include a higher proportion of mildly deleterious mutations that have not yet been removed through purifying selection.

Determining when ZIKV arrived in specific regions helps to elucidate the spread of the outbreak and track rising incidence of possible complications of ZIKV infection. The majority of the ZIKV genomes from our study fall into four major clades from different geographic regions, for which we estimated a likely date for ZIKV arrival. In each case, the date was months earlier than the first confirmed, locally transmitted case, indicating ongoing local circulation of ZIKV before its detection. In Puerto Rico, the estimated date was 4.5 months earlier than the first confirmed local case [3]; it was 8 months earlier in Honduras [2], 5.5 months earlier in Colombia [24], and 9 months earlier for the Caribbean-continental U.S. clade [4]. In each case, the arrival date represents the estimated time to the most recent common ancestor (tMRCA) for the corresponding clade in our phylogeny (Figure 4.2C; see Figure D.3 and Table D.2 for details). Similar temporal gaps between the tMRCA of local transmission chains and the earliest detected cases were seen when chikungunya virus emerged in the Americas [25]. We also observed evidence for several in-

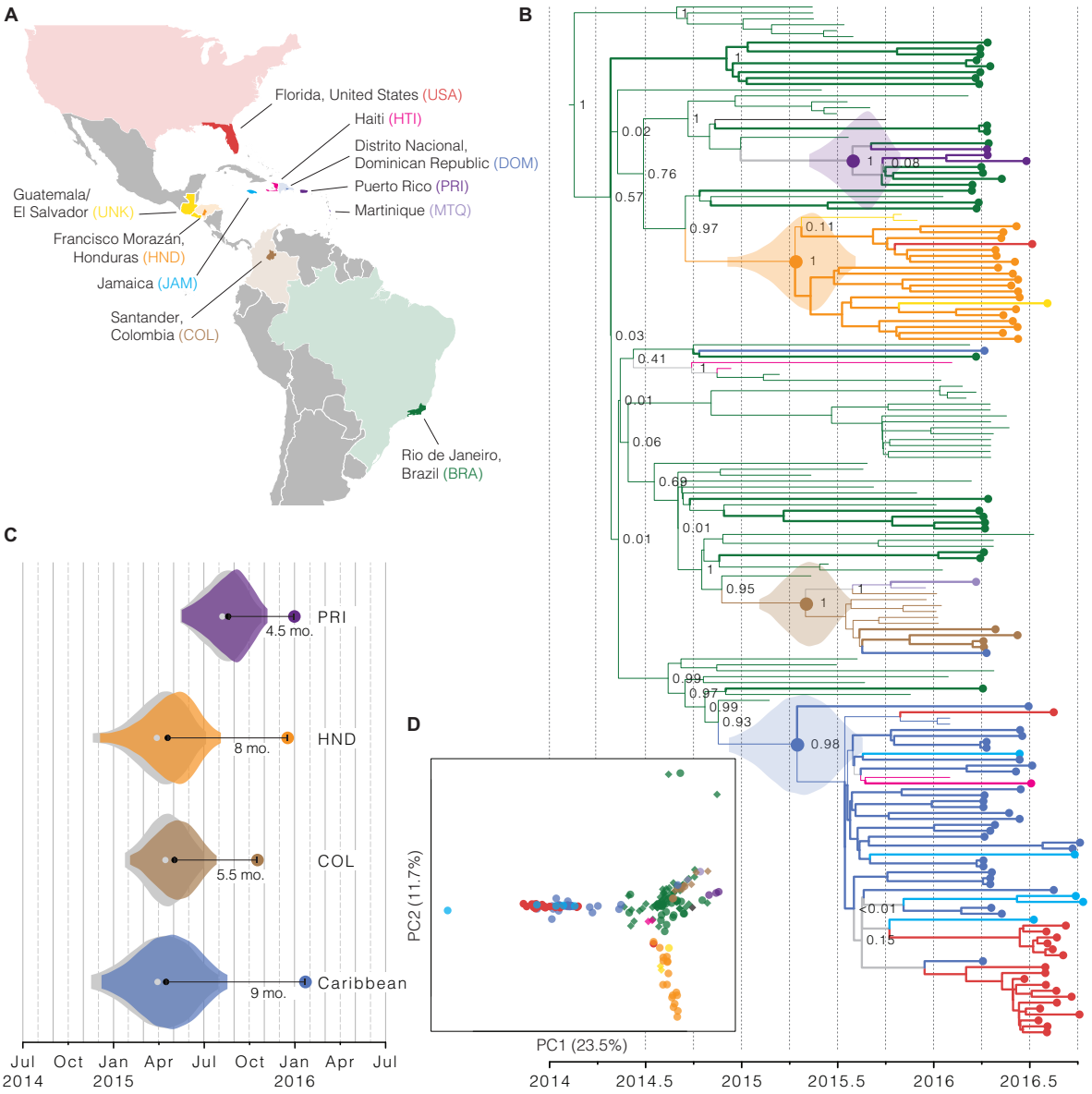


Figure 4.2: Zika virus spread throughout the Americas. (**A**) Samples were collected in each of the colored countries or territories. Specific state, department, or province of origin for samples in this study is labeled if known. (**B**) Maximum clade credibility tree. Dotted tips, genomes generated in this study. Node labels are posterior probabilities indicating support for the node. Violin plots denote probability distributions for the tMRCA of four highlighted clades. (**C**) Time elapsed between estimated tMRCA and date of first confirmed, locally transmitted case. Color, distributions based on relaxed clock model (also shown in B); grey, strict clock. Caribbean clade includes the continental United States. (**D**) Principal components analysis of variants. Circles, data generated in this study; diamonds, other publicly available genomes from this outbreak. Percentage of variance explained by each component is indicated on axis.

troductions of ZIKV into the continental United States, and found that sequences from mosquito and human samples collected in Florida cluster together, consistent with the finding of local ZIKV transmission in Florida in Grubaugh et al. [10].

Principal components analysis (PCA) is consistent with the phylogenetic observations (Figure 4.2D). It shows tight clustering among ZIKV genomes from the continental United States, the Dominican Republic, and Jamaica. ZIKV genomes from Brazil and Colombia are similar and distinct from genomes sampled in other countries. ZIKV genomes from Honduras form a third cluster that also contains genomes from Guatemala or El Salvador. The PCA results show no clear stratification of ZIKV within Brazil.

4.3.3 Zika Virus Variants with Possible Biological Significance

Genetic variation can provide important insights into ZIKV biology and pathogenesis and can reveal potentially functional changes in the virus. We observed 1,030 mutations in the complete dataset, and they were well distributed across the genome (Figure 4.3A). Any effect of these mutations cannot be determined from these data; however, the most likely candidates for functional mutations would be among the 202 nonsynonymous mutations and the 32 mutations in the 5' and 3' untranslated regions (UTRs). Adaptive mutations are more likely to be found at high frequency or to be seen multiple times, although both effects can also occur by chance. We observed five positions with nonsynonymous mutations at more than 5% minor allele frequency that occurred on two or more branches of the tree (Figure 4.3B); two of these (at positions 4,287 and 8,991) occurred together and might represent incorrect placement of a Brazil branch in the tree. The remaining

three are more likely to represent multiple nonsynonymous mutations; one (at 9,240) appears to involve nonsynonymous mutations to two different alleles.

To assess the possible biological significance of these mutations, we looked for evidence of selection in the ZIKV genome. Viral surface glycoproteins are known targets of positive selection, and mutations in these proteins can confer adaptation to new vectors [32] or aid immune escape [33, 34]. We therefore searched for an excess of nonsynonymous mutations in the ZIKV envelope glycoprotein (E). However, the nonsynonymous substitution rate in E proved to be similar to that in the rest of the coding region (Figure 4.3C, left); moreover, amino acid changes were significantly more conservative in that region than elsewhere (Figure 4.3C, middle and right). Any diversifying selection occurring in the surface protein thus appears to be operating under selective constraint. We also found evidence for purifying selection in the ZIKV 3' UTR (Figure 4.3D), which is important for viral replication [35].

While the transition-to-transversion ratio (6.98) was within the range seen in other viruses [36], we observed a considerably higher frequency of C-to-T and T-to-C substitutions than other transitions (Figures 4.3D, D.4). This enrichment was apparent both in the genome as a whole and at fourfold degenerate sites, where selection pressure is minimal. Many processes could contribute to this conspicuous mutation pattern, including mutational bias of the ZIKV RNA-dependent RNA polymerase, host RNA editing enzymes (for example, APOBECs, ADARs) acting upon viral RNA, and chemical deamination, but further investigation is required to determine the cause of this phenomenon.

Mismatches between PCR assays and viral sequence are a potential source of poor diagnostic

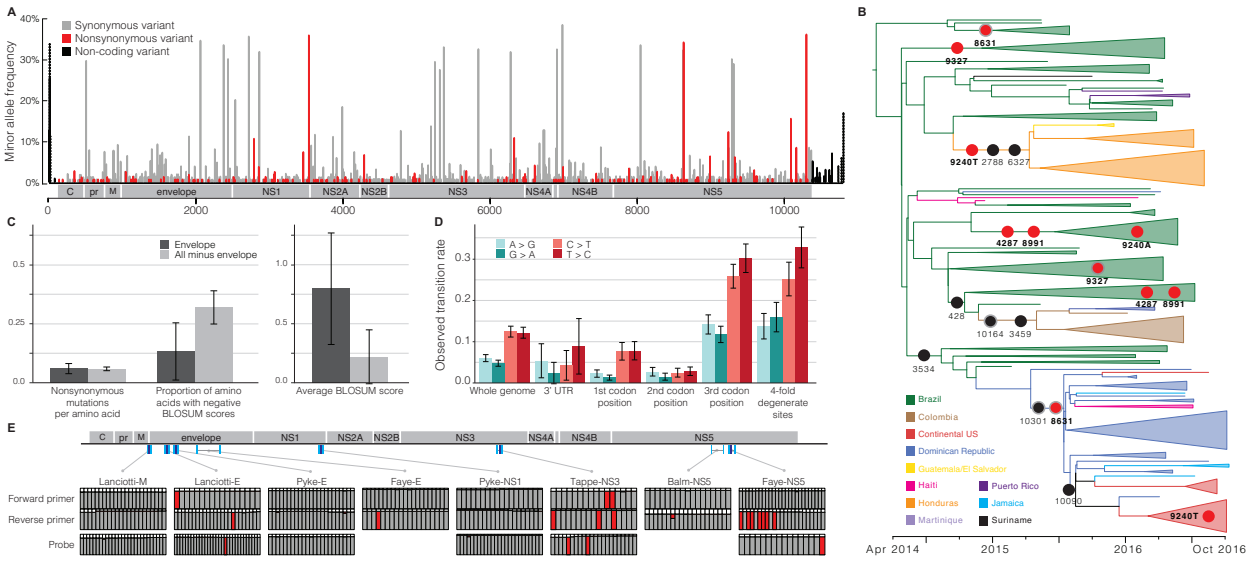


Figure 4.3: Geographic and genomic distribution of Zika virus variation. (**A**) Location of variants in the ZIKV genome. The minor allele frequency is the proportion of the 174 genomes from this outbreak that share a variant. Dotted bars, <25% of samples had a base call at that position. (**B**) Phylogenetic distribution of nonsynonymous variants with minor allele frequency >5%, shown on the branch where the mutation is most likely to have occurred. Grey outline, variant might be on next-most ancestral branch (in two cases, two branches upstream), but exact location is unclear because of missing data. Red circles, variants occurring at more than one location in the tree. (**C**) Conservation of the ZIKV envelope (E) region. Left, nonsynonymous variants per amino acid for the E region (dark grey) and the rest of the coding region (light grey). Middle, proportion of nonsynonymous variants resulting in negative BLOSUM62 scores, which indicate unlikely or extreme substitutions (p<0.039, χ^2 test). Right, average of BLOSUM62 scores for nonsynonymous variants (p<0.037, two-sample t-test). (**D**) Constraint in the ZIKV 3' UTR and observed transition rates over the ZIKV genome. (**E**) ZIKV diversity in diagnostic primer and probe regions. Top, locations of published probes (dark blue) and primers (cyan)[26–31] on the ZIKV genome. Bottom, each column represents a nucleotide position in the probe or primer. Colors in the column indicate the fraction of ZIKV genomes (out of 174) that matched the probe/primer sequence (grey), differed from it (red), or had no data at that position (white).

performance in this outbreak [37]. To assess the potential influence of ongoing viral evolution on diagnostic function, we compared eight published qRT-PCR-based primer/probe sets to our data. We found numerous sites at which the probe or primer did not match an allele found among the 174 ZIKV genomes from the current dataset (Figure 4.3E). In most cases, the discordant allele was shared by all outbreak samples, presumably because it was present in the Asian lineage that entered the Americas. These mismatches could affect all uses of the diagnostic assay in the outbreak. We also found mismatches from new mutations that occurred after ZIKV entry into the Americas. Most of these were present in less than 10% of samples, although one was seen in 29%. These observations suggest that genome evolution has not caused widespread degradation of diagnostic performance during the course of the outbreak, but that mutations continue to accumulate and ongoing monitoring is needed.

4.3.4 Within-Host Variant Detection in Low-Titer Viruses

Viral within-host variants can be useful for understanding transmission between hosts, and for elucidating viral dynamics and evolutionary processes. This variation comes from the presence of viral quasispecies [38, 39], viruses that are very similar but not identical within a single host. Quasispecies are particularly common in RNA viruses, which have high mutation rates, short generation times, and large population sizes [40, 41]. The coexistence of multiple variants may be advantageous to viruses, and provide opportunities for immune and vaccine escape [42, 43]. Therefore, studying these variants may reveal functionally-important mutations, while also providing clues about within-host dynamics and viral pathogenesis [44]. If multiple viral quasispecies are passed

between hosts during an infection event [45–47], they may also be useful in understanding transmission patterns, especially if between-host viral variation is otherwise limited.

Given the potential of within-host variation to inform transmission and biology, we attempted to identify intrahost variants (iSNVs) in our ZIKV samples. Since we had sequencing data from two independent sequencing methods (amplicon sequencing and hybrid capture), we also wanted to test concordance between these methods and potentially improve upon our method [45] for filtering out low-frequency iSNVs likely due to contamination or sequencing error.

We called iSNVs on genomes generated by each sequencing method. We initially used the frequency filters established in Gire et al. [45], which require an iSNV to be identified in five forward and reverse reads and limits strand bias (the ratio between reads with the variant on the two strands) to 10×. Using this method, we quickly noticed that very few iSNVs were identified in samples generated by the amplicon sequencing method. Given the substantial differences in sequencing preparation methods, we hypothesized that a different set of requirements may be necessary to accurately filter data generated using this method. Therefore, we removed all filters from both methods and compared the resulting variants, assuming variants generated by hybrid capture and passing the Gire et al. filters to be correct (we call these 'verified' iSNVs). We additionally required a minimum read depth at each iSNV position, with the aim of eliminating difference in coverage as a reason for unmatched calls (see Chapter 2, Figure B.3).

The results of this analysis, summarized in Figure 4.4, show that high-frequency (>20%) iSNVs are consistently identified by both methods. At lower frequencies, however, the amplicon sequencing method misses variants that we accept as true (Figure 4.4A and Table D.3A, which shows

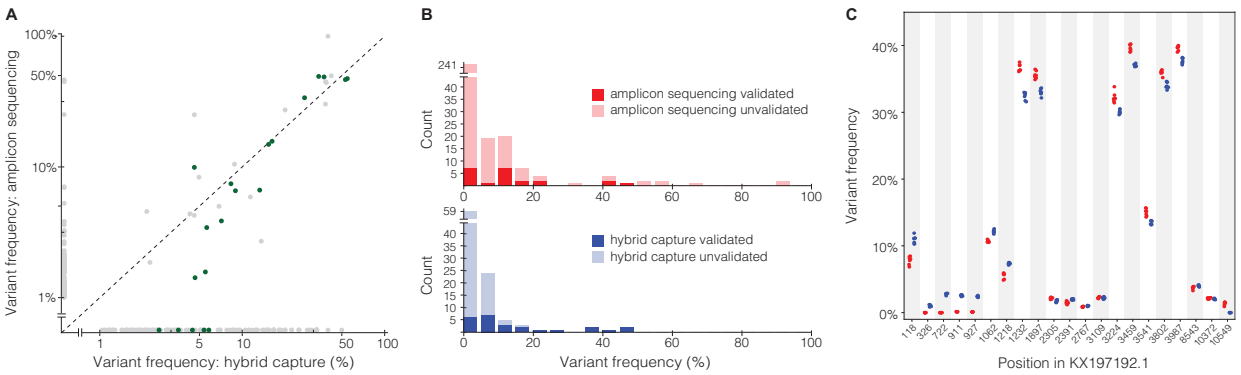


Figure 4.4: Comparison of within-host variants across methods and replicates. (**A**) Within-sample variant frequencies across methods. Each point is a variant in a clinical or mosquito sample and is plotted on a log-log scale. Green points, 'verified' variants detected by hybrid capture that pass strand bias and frequency filters. Frequencies <1% are shown at 0%. (**B**) Counts of within-sample variants across two technical replicates for each method. Variants are plotted in the frequency bin corresponding to the higher of the two detected frequencies. (**C**) Within-sample variants for a single cultured isolate (PE243) across seven technical replicates. Each point is a variant in a replicate identified using amplicon sequencing (red) or hybrid capture (blue). Variants are plotted if the pooled frequency across replicates by either method is $\geq$1%.

that 25% of verified iSNVs are not identified using amplicon sequencing). Additionally, we find that amplicon sequencing identifies a large number of presumably spurious mutations not identified in hybrid capture replicates, or even in a second amplicon sequencing replicate (Table D.3).

This investigation also demonstrates the importance of filtering within-host variant calls, regardless of method. Low levels of contamination and sequencing error are common contributors to low-frequency variants; these issues are unavoidable with current technologies and must be addressed by computational filtering. Figures 4.4B and Table D.3B show that nearly 75% of variants from hybrid capture sequences failed to replicate, but that this problem disappears when implementing a strand bias filter. However, the low number of variants passing this filter (n=8) suggests that this method may be too conservative, that the samples were not sequenced to high enough read depth for iSNV identification, or that ZIKV may lack substantial within-host variation. Opti-

mizing methods and further testing on ZIKV will be required to better understand this result and potentially generate iSNVs useful for studying ZIKV evolution or transmission. Identifying thresholds that successfully filter out spurious variants will be especially useful for amplicon sequencing, and will hopefully generate reliable iSNVs from this method. Notably, in a cultured ZIKV sample with high viral content, both methods produced the same results and were internally consistent at nearly all frequencies (Figure 4.4C), again suggesting that issues in iSNV identification may be exacerbated by enrichment of low quality samples.

Although this analysis did not produce enough validated variants for ZIKV transmission analysis, it clearly demonstrates the need for appropriate filtering of variants, and cautions against drawing conclusions from within-host variants without replication and validation. It also adds to our understanding of amplicon and hybrid capture sequencing, both of which are still widely used.

4.4 Discussion

Sequencing low-titer viruses such as ZIKV directly from clinical samples presents several challenges that are likely to have contributed to the paucity of genomes available from the current outbreak. While the development of technical and analytical methods will surely continue, we note that factors upstream in the process, including collection site and cohort, were strong predictors of sequencing success in our study (Figure D.1). This finding highlights the importance of continuing development and implementation of best practices for sample handling, without disrupting standard clinical workflows, for wider adoption of genome surveillance during outbreaks. Additional

sequencing, however challenging, remains critical to ongoing investigation of ZIKV biology and pathogenesis. Together with Faria et al. [9] and Grubaugh et al. [10], this study advances both technological and collaborative strategies for genome surveillance in the face of unexpected outbreak challenges.

4.5 Methods

4.5.1 Ethics Statement

The clinical studies from which samples were obtained were evaluated and approved by the relevant Institutional Review Boards/Ethics Review Committees at Hospital General de la Plaza de la Salud (Santo Domingo, Dominican Republic), University of the West Indies (Kingston, Jamaica), Universidad Nacional Autónoma de Honduras (Tegucigalpa, Honduras), Oswaldo Cruz Foundation (Rio de Janeiro, Brazil), Centro de Investigaciones Epidemiologicas–Universidad Industrial de Santander (Bucaramanga, Colombia), Massachusetts Department of Public Health (Jamaica Plain, Massachusetts), and Florida Department of Health (Tallahassee, Florida). Informed consent was obtained from all participants enrolled in studies at Hospital General de la Plaza de la Salud, Universidad Nacional Autónoma de Honduras, Oswaldo Cruz Foundation, and Universidad Industrial de Santander. IRBs at the University of West Indies, Massachusetts Department of Public Health, and Florida Department of Health granted waivers of consent given this research with leftover clinical diagnostic samples involved no more than minimal risk. Harvard University and Massachusetts Institute of Technology (MIT) Institutional Review Boards/Ethics Review Committees provided

approval for sequencing and secondary analysis of samples collected by the aforementioned institutions.

4.5.2 Sample collections and study subjects

Patients with suspected ZIKV infection (including high-risk travellers) were enrolled through study protocols at multiple aforementioned collection sites. Clinical samples (including blood, urine, cerebrospinal fluid, and saliva) were obtained from suspected or confirmed ZIKV cases and from high-risk travellers.

4.5.3 Viral RNA Isolation

RNA was isolated following the manufacturer's standard operating protocol for 0.14–1 ml samples [8] using the QIAamp Viral RNA Minikit (Qiagen), except that in some cases 0.1-M final concentration of β-mercaptoethanol (as a reducing agent) or 40 μg/ml final concentration of linear acrylamide (Ambion) (as a carrier) were added to AVL buffer before inactivation. Extracted RNA was resuspended in AVE buffer or nuclease-free water. In some cases, viral samples were concentrated using Vivaspin-500 centrifugal concentrators (Sigma-Aldrich) before inactivation and extraction. In these cases, 0.84 ml of sample was concentrated to 0.14 ml by passing through a 30 kDa filter and discarding the flow-through.

4.5.4 Carrier RNA and Host rRNA Depletion

In a subset of human samples, carrier poly(rA) RNA and host rRNA were depleted from RNA samples using RNase H selective depletion [22, 48]. In brief, oligo d(T) (40 nt long) and/or DNA probes complementary to human rRNA were hybridized to the sample RNA. The sample was then treated with 15 units Hybridase (Epicentre) for 30 min at 45°C. The complementary DNA probes were removed by treating each reaction with an RNase-free DNase (Qiagen) according to the manufacturer's protocol. Following depletion, samples were purified using 1.8× volume AMPure RNAclean beads (Beckman Coulter Genomics) and eluted into 10 μl water for cDNA synthesis.

4.5.5 Illumina Library Construction and Sequencing

cDNA synthesis was performed as described in previously published RNA-seq methods [22]. To track potential cross-contamination, 50 fg synthetic RNA (gift from M. Salit, NIST) was spiked into samples using unique RNA for each individual ZIKV sample. ZIKV negative control cDNA libraries were prepared from water, human K-562 total RNA (Ambion), or EBOV (GenBank accession: KY425633.1) seed stock; ZIKV positive controls were prepared from ZIKV Senegal (isolate HD78788) or ZIKV Pernambuco (isolate PE243; GenBank accession: KX197192.1) seed stock. The dual index Accel-NGS 2S Plus DNA Library Kit (Swift Biosciences) was used for library preparation. Approximately half of the cDNA product was used for library construction, and indexed libraries were generated using 18 cycles of PCR. Each individual sample was indexed with a unique barcode. Libraries were pooled at equal molarity and sequenced on the Illumina HiSeq 2500 or

MiSeq (paired-end reads) platforms.

4.5.6 Amplicon-based cDNA Synthesis and Library Construction

ZIKV amplicons were prepared as described [10, 21], similarly to 'RNA jackhammering' for preparing low-input viral samples for sequencing [49], with slight modifications. After PCR amplification, each amplicon pool was quantified on a 2200 Tapestation (Agilent Technologies) using High Sensitivity D1000 ScreenTape (Agilent Technologies). 2 μl of a 1:10 dilution of the amplicon cDNA was loaded and the concentration of the 350–550 bp fragments was calculated. The cDNA concentration, as reported by the Tapestation, was highly predictive of sequencing outcome (that is, whether a sample passed genome assembly thresholds) (Figure D.5). cDNA from each of the two amplicon pools was mixed equally (10–25 ng each) and libraries were prepared using the dual index Accel-NGS 2S Plus DNA Library Kit (Swift Biosciences) according to the manufacturer's protocol. Libraries were indexed with a unique barcode using seven cycles of PCR, pooled equally and sequenced on the Illumina MiSeq (250 bp paired-end reads) platform. Primer sequences were removed by hard trimming the first 30 bases for each insert read before analysis.

4.5.7 Zika Virus Hybrid Capture

Virus hybrid capture was performed as previously described [22]. Probes were created to target ZIKV and chikungunya virus (CHIKV). Candidate probes were created by tiling across publicly available sequences for ZIKV and CHIKV on NCBI GenBank [50]. Probes were selected from among these candidate probes to minimize the number used while maintaining coverage of the observed

diversity of the viruses. Alternating universal adapters were added to allow two separate PCR amplifications, each consisting of non-overlapping probes.

The probes were synthesized on a 12k array (CustomArray). The synthesized oligos were amplified by two separate emulsion PCR reactions with primers containing T7 RNA polymerase promoter. Biotinylated baits were in vitro transcribed (MEGAshortscript, Ambion) and added to prepared ZIKV libraries. The baits and libraries were hybridized overnight (16 h), captured on streptavidin beads, washed, and re-amplified by PCR using the Illumina adaptor sequences. Capture libraries were then pooled and sequenced. In some cases, a second round of hybrid capture was performed on PCR-amplified capture libraries to further enrich the ZIKV content of sequencing libraries (Figure D.6). In the main text, 'hybrid capture' refers to a combination of hybrid capture sequencing data and data from the same libraries without capture (unbiased), unless explicitly distinguished.

4.5.8 Genome Assembly

We assembled reads from all sequencing methods into genomes using viral-ngs v1.13.3 [46, 51]. We taxonomically filtered reads from amplicon sequencing against a ZIKV reference, GenBank accession: KU321639.1. To compute results on individual replicates, we de novo assembled these and scaffolded against KU321639.1. To obtain final genomes for analysis, we pooled data from multiple replicates of a sample, de novo assembled, and scaffolded against KX197192.1. For all assemblies, we set the viral-ngs 'assembly_min_length_fraction_of_reference' and 'assembly_min_unambig' parameters to 0.01. For amplicon sequencing data, unambiguous base calls required at least 90%

of reads to agree in order to call that allele ('major_cutoff'=0.9); for hybrid capture data, we used the default threshold of 50%. We modified viral-ngs so that calls to GATK's UnifiedGenotyper set 'min_indel_count_for_genotyping' to 2.

At three sites with insertions or deletions (indels) in the consensus genome coding region (CDS), we corrected the genome using Sanger sequencing of the RT-PCR product (namely, at 3,447 in the genome for sample DOM_2016_BB-0085-SER; at 5,469 in BRA_2016_FC-DQ12D1-PLA; and at 6,516–6,564 in BRA_2016_FC-DQ107D1-URI, coordinates as in KX197192.1). At other indels in the consensus genome CDS, we replaced the indel with ambiguity.

Depth-of-coverage values from amplicon sequencing include read duplicates. In all other cases, we removed duplicates with viral-ngs.

4.5.9 Identification of Non-Zika Viruses in Samples

Using Kraken v0.10.6 [52] in viral-ngs, we built a database that included its default full database (which incorporates all bacterial and viral whole genomes from RefSeq [53] as of October 2015). Additionally, we included the whole human genome (hg38), genomes from PlasmoDB [54], sequences covering mosquito genomes (Aedes aegypti, Aedes albopictus, Anopheles albimanus, Anopheles quadrimaculatus, Culex quinquefasciatus, and the outgroup Drosophila melanogaster) from GenBank [50], protozoa and fungi whole genomes from RefSeq, SILVA LTP 16S rRNA sequences [55], and all sequences from NCBI's viral accession list [56] (as of October 2015) for viral taxa that have human as a host.

For each sample, we ran Kraken on data from unbiased sequencing replicates (not including

hybrid capture data) and searched its output reports for viral taxa with more than 100 reported reads. We manually filtered the results, removing ZIKV, bacteriophages, and known laboratory contaminants. For each sample and its associated taxa, we assembled genomes using viral-ngs as described above; the results are in Table D.1. We used the following genomes for taxonomically filtering reads and as the reference for assembly: KJ741267.1 (cell fusing agent virus), AY292384.1 (deformed wing virus), NC_001477.1 (dengue virus type 1) and LC164349.1 (JC polyomavirus). When reporting sequence identity of an assembly to its taxon, we used BLASTN43 to determine the identity between the sequence and the reference used for its assembly.

To focus on metagenomics of mosquito pools (Table D.1B), we considered unbiased sequencing data from eight mosquito pools (not including hybrid capture data). We first ran the depletion pipeline of viral-ngs on raw data and then ran the viral-ngs Trinity [57] assembly pipeline on the depleted reads to assemble them into contigs. We pooled contigs from all mosquito pool samples and identified all duplicate contigs with sequence identity >95% using CD-HIT [58]. Additionally, we used predicted coding sequences from Prodigal 2.6.3 [59] to identify duplicate protein sequences at >95% identity. We classified contigs using BLASTN [60] against nt and BLASTX [60] against nr (as of February 2017) and discarded all contigs with an E value greater than 1×10^{-4}. We define viral contigs as contigs that hit a viral sequence, and we manually removed all reverse-transcriptase-like contigs owing to their similarity to retrotransposon elements within the Aedes aegypti genome. We categorized viral contigs with less than 80% amino acid identity to their best hit as likely novel viral contigs.

4.5.10 Relationship Between Metadata and Sequencing Outcome

To determine whether available sample metadata are predictive of sequencing outcome, we tested the following variables: sample collection site, patient gender, patient age, sample type, and the number of days between symptom onset and sample collection (collection interval). To describe sequencing outcome of a sample S, we used the following response variable Y_S: mean ($\{I(R) * (\text{number of unambiguous bases in } R)$ for all amplicon sequencing replicates R of $S\}$), where $I(R) = 1$ if median depth of coverage of $R \geq 275$ and $I(R) = 0$ otherwise.

We excluded the saliva, cerebrospinal fluid, and whole blood sample types owing to sample number (n=1), and also excluded mosquito pool samples and rows with missing values. We excluded samples from one collection site (prefix JAM_2016_WI-) because most had missing values. We treated samples with type 'Plasma EDTA' as having type 'Plasma'. We treated the collection interval variable as categorical (0–1, 2–3, 4–6, and 7+ days).

With a single model we underfit the zero counts, possibly because many zeros (samples without a replicate that passed ZIKV assembly) are truly ZIKV-negative. We thus view the data as coming from two processes: one determining whether a sample is ZIKV-positive or ZIKV-negative, and another that determines, among the observed passing samples, how much of a ZIKV genome we are able to sequence. We modelled the first process, predicting whether a sample is passing, with logistic regression (in R using GLM [61] with binomial family and logit link); here, the observed passing samples are the samples S for which $Y_S \geq 2,500$. For the second, we performed a beta regression, using only the observed passing samples, of Y_S divided by ZIKV genome length on the

predictor variables. We implemented this in R using the betareg package [62] and transformed fractions from the closed unit interval to the open unit interval as the authors suggest.

To test the significance of predictor variables, we used a likelihood ratio test. For variable X_i we compared a full model (with all predictors) against a model that used all predictors except X_i. The results of these tests are shown in Figure D.1A,D. We explored the effects of sample type and collection interval on obtaining a passing assembly in Figure D.1B and C, respectively. Error bars are 95% confidence intervals derived from binomial distributions. We explored the effects of these same two variables on Y_S (in passing samples only) in Figure D.1E–F.

4.5.11 Criteria for Pooling Across Replicates

We attempted to sequence one or more replicates of each sample and attempted to assemble a genome from each replicate. We discarded data from any replicates whose assembly showed high sequence similarity, in any part of the genome, to our assembly of the genome in a sample consisting of an African (Senegal) lineage (strain HD78788) of ZIKV. We used this sample as a positive control throughout this study, and considered its presence in the assembly of a clinical or mosquito pool sample to be evidence of contamination. Similarly, we discarded data from four replicates belonging to samples from the Dominican Republic because they yielded assemblies that were unexpectedly identical or highly similar to our assembly of the ZIKV isolate PE243 genome, another positive control used in this study. We also discarded data from replicates that showed evidence of contamination, at the RNA stage, by the baits used in hybrid capture; we detected these by looking for adapters that were added to these probes for amplification.

For amplicon sequencing, we considered an assembly of a replicate to be 'passing' if it contained at least 2,500 unambiguous base calls and had a median depth of coverage of at least 275× over its unambiguous bases (depth includes duplicate reads). For the unbiased and hybrid capture approaches, we considered an assembly of a replicate 'passing' if it contained at least 4,000 unambiguous base calls. For each approach, the unambiguous base threshold was based on an observed density of negative controls below the threshold (Figure 4.1A). For amplicon sequencing assemblies, we added a coverage depth threshold because coverage depth was roughly binary across replicates, with negative controls falling in the lower class. On the basis of these thresholds, zero of 99 negative controls used throughout our sequencing runs yielded passing assemblies and 32 of 32 positive controls yielded passing assemblies.

We considered a sample to have a passing assembly if any of its replicates, by either method, yielded an assembly that passed the above thresholds. For each sample with at least one passing assembly, we pooled read data across replicates for each sample, including replicates with assemblies that did not pass the assembly thresholds. When data were available from both amplicon sequencing and unbiased/hybrid capture approaches, we pooled amplicon sequencing data separately from data produced by the unbiased and hybrid capture approaches, the latter two of which were pooled together (henceforth, the 'hybrid capture' pool). We then assembled a genome from each set of pooled data. When assemblies on pooled data were available from both approaches, we selected for downstream analysis the assembly from the hybrid capture approach if it had at least 10,267 unambiguous base calls (95% of the reference genome used, GenBank accession: KX197192.1); when this condition was not met, we selected the one that had more unambiguous base calls.

The number of ZIKV genomes publicly available before this study was the result of an NCBI GenBank [50] search for ZIKV in February 2017. We filtered any sequences with length <4,000 nt, excluded sequences that are being published as part of this study or in Faria et al. [9] or Grubaugh et al. [10], excluded sequences from non-human hosts, and excluded sequences labelled as having been passaged. We counted fewer than 100 sequences, the precise number depending on details of the count.

4.5.12 Visualization of Coverage Depth Across Genomes

For amplicon sequencing data, we plotted coverage across the 110 samples that yielded a passing assembly by amplicon sequencing (Figure 4.1B). With viral-ngs, we aligned depleted reads to the reference sequence KX197192.1 using the novoalign aligner with options '-r Random -l 40 -g 40 -x 20 -t 100 -k'. Because of the nature of amplicon sequencing, duplicates were not identified or removed. We binarized depth at each nucleotide position, showing red if depth of coverage was at least 100×. Rows (samples) are hierarchically clustered to ease visualization.

For hybrid capture sequencing data, we plotted depth of coverage across the 37 samples that yielded a passing assembly (Figure 4.1C). We aligned reads as described above for amplicon sequencing data, except we removed duplicates. For each sample, we calculated the depth of coverage at each nucleotide position. We then scaled the values for each sample so that each would have a mean depth of 1.0. At each nucleotide position, we calculated the median depth across the samples, as well as the 20th and 80th percentiles. We plotted the mean of each of these metrics within a 200-nt sliding window.

4.5.13 Multiple Sequence Alignments

We aligned ZIKV consensus genomes using MAFFT v7.221 [63] with the following parameters: '--maxiterate 1000 --ep 0.123 --localpair'.

4.5.14 Analysis of Within- and Between-Sample Variants

To measure overall per-base discordance between consensus genomes produced by amplicon sequencing and hybrid capture, we considered all sites at which base calls were made in both the amplicon sequencing and hybrid capture consensus genomes of a sample, and we calculated the fraction in which the bases were not in agreement. To measure discordance at polymorphic sites, we searched for positions with a polymorphism in all genomes generated in this study that we selected for downstream analysis (see Section 4.5.11 for choosing among the amplicon sequencing and hybrid capture genome when both are available). We then looked at these positions in genomes that were available from both methods, and we calculated the fraction in which the alleles were not in agreement.

To measure discordance at minor alleles, we searched for minor alleles in all genomes generated in this study that we selected for downstream analysis. We then looked at all sites at which there was a minor allele and for which genomes from both methods were available, and we calculated the fraction in which the alleles were not in agreement. For these calculations, we tolerated partial ambiguity (for example, 'Y' is concordant with 'T'). If one genome had full ambiguity ('N') at a position and the other genome had an indel, we counted the site as discordant; otherwise, if one

genome had full ambiguity, we did not count the site.

After assembling genomes, we identified within-sample variants by running V-Phaser 2.0 via viral-ngs [51] on all pooled reads mapping to each sample assembly. When determining per-library allele counts at each variant position, we modified viral-ngs to require a minimum base (Phred) quality score of 30 for all bases, discard anomalous read pairs, and use per-base alignment quality (BAQ) in its calls to SAMtools [64] mpileup. This is particularly helpful for filtering spurious amplicon sequencing variants because all generated reads start and end at a limited number of positions (owing to the pre-determined tiling of amplicons across the genome). Because amplicon sequencing libraries were sequenced using 250 bp paired-end reads, bases near the middle of the approx. 450 nucleotide amplicons fall at the end of both paired reads, where quality scores drop and incorrect base calls are more likely. To determine the overall frequency of each variant in a sample, we summed allele counts (calculated using SAMtools [64] mpileup via viral-ngs) across libraries.

When comparing variant frequencies between amplicon sequencing (seven technical replicates) and hybrid capture (seven technical replicates) replicates of the PE243 positive control (Figure 4.4C), we included only positions at which the mean (pooled) frequency across replicates within at least one method was $\geq$1%. When comparing allele frequencies between replicate libraries, we restricted the sample set to only samples with a passing assembly in both methods, and included only samples with two or more replicates. By contrast, when comparing alleles across methods, we included samples that have a passing assembly by either method, with any number of replicates. For these comparisons, we included only positions with a minor variant; that is, positions for which both libraries/methods had an allele at 100% were removed, even if the single allele

differed between the two libraries/methods. Additionally, we considered any allele with frequency <1% as not found (0%).

When comparing allele frequencies across methods: let f_a and f_{hc} be frequencies in amplicon sequencing and hybrid capture, respectively. If both are non-zero, we included an allele only if the read depth at its position was $\geq 1/min(f_a, f_{hc})$ in both methods, and if depth at the position was at least 100× for hybrid capture and 275× for amplicon sequencing. If $f_a = 0$, we required a read depth of $max(1/f_{hc}, 275)$ at the position in the amplicon sequencing method; similarly, if $f_{hc} = 0$ we required a read depth of $max(1/f_a, 100)$ at the position in the hybrid capture method. This was to eliminate lack of coverage as a reason for discrepancy between two methods. When comparing allele frequencies across sequencing replicates within a method, we imposed only a minimum read depth (275× for amplicon sequencing and 100× for hybrid capture), but required this depth in both libraries. In samples with more than two replicates, we considered only the two replicates with the highest depth at each variant position.

We considered allele frequencies from hybrid capture sequencing 'verified' if they passed the strand bias and frequency filters described in Gire et al. [45], with the exception that we imposed a minimum allele frequency of 1% and allowed a variant identified in only one library if its frequency was ≥5%. In Figure 4.4B and Table D.3, we considered variants 'validated' if they were present at ≥1% frequency in both libraries or methods. When comparing two libraries for a given method *M* (amplicon sequencing or hybrid capture): the proportion unvalidated is the fraction, among all variants in *M* at ≥1% frequency in at least one library, of the variants that are at ≥1% frequency in exactly one of the two libraries. Similarly, when comparing methods: the proportion unvalidated

for a method M is the fraction, among all variants at $\geq$1% frequency in M, of the variants that are at $\geq$1% frequency in M and <1% frequency in the other method.

We called SNPs on the aligned genomes using Geneious version 9.1.7 [65]. We converted all fully or partially ambiguous calls, which are treated by Geneious as variants, into missing data. We then removed all sites that were no longer polymorphic from the SNP set and re-calculated allele frequencies. A nonsynonymous mutation is shown on the tree (Figure 4.3B) if it includes an allele that is nonsynonymous relative to the ancestral state (see Section 4.5.16 below) and has a minor allele frequency of >5%; all occurrences of nonsynonymous alleles are shown. (Two mutations, at positions 2,853 and 7,229, had nominal derived allele frequencies over 95%; in both cases, the ancestral allele was seen only in a small clade within the tree, suggesting that the ancestral allele was incorrectly assigned. These are not shown.) We placed mutations at a node such that the node leads only to samples with the mutation or with no call at that site. Uncertainty in placement occurs when a sample lacks a base call for the corresponding mutation; in this case, we placed the mutation on the most recent branch for which we have available data. We also used this ancestral ZIKV state to count the frequency of each type of substitution over various regions of the ZIKV genome, per number of available bases in each region (Figure 4.3D).

We quantified the effect of nonsynonymous mutations using the original BLOSUM62 scoring matrix for amino acids [66], in which positive scores indicate conservative amino acid changes and negative scores unlikely or extreme substitutions. We assessed statistical significance for equality of proportions by χ^2 test (Figure 4.3C, middle), and for difference of means by two-sample t-test with Welch-Satterthwaite approximation of d.f. (Figure 4.3C, right). Error bars are 95% confidence

intervals derived from binomial distributions (Figure 4.3C, left and middle; Figure 4.3D) or Student's *t* distributions (Figure 4.3C, right).

4.5.15 Maximum likelihood estimation and root-to-tip regression

We generated a maximum likelihood tree using a multiple sequence alignment that included genomes generated in this study, as well as a selection of other available sequences from the Americas, Southeast Asia, and the Pacific. We ran PhyML [67] with the GTR substitution model and four gamma substitution rate categories; for the tree search operation, we used 'BEST' (best of NNI and SPR). In FigTree v1.4.2 [68], we rooted the tree on the oldest sequence used as input (GenBank accession: EU545988.1).

We used TempEst v1.5 [69], which selects the best-fitting root with a residual mean squared function, to estimate root-to-tip distances. We performed regression in R with the lm function [61] of distances on dates. The relationship between root-to-tip divergence and sample dates (Figure D.2) supports the use of a molecular clock analysis in this study.

4.5.16 Molecular Clock Phylogenetics

For molecular clock phylogenetics, we made a multiple sequence alignment from the genomes generated in this study combined with a selection of other available sequences from the Americas. We did not use sequences from outside the outbreak in the Americas. Among ZIKV genomes published and publicly available on NCBI GenBank [50], we selected 32 from the Americas that had at least 7,000 unambiguous bases, were not labelled as having been passaged more than once, and

had location metadata. We also used 32 genomes from Brazil published in Faria et al. [9] that met the same criteria.

We used BEAST v1.8.4 to perform molecular clock analyses [70]. We used sampled tip dates to handle inexact dates [71]. Because of sparse data in non-coding regions, we used only the CDS as input. We used the SRD06 substitution model on the CDS, which uses HKY with gamma site heterogeneity and partitions codons into two partitions (positions (1+2) and 3) [72]. To perform model selection, we tested three coalescent tree priors: a constant-size population, an exponential growth population, and a Bayesian Skyline tree prior (ten groups, piecewise-constant model) [73]. For each tree prior, we tested two clock models: a strict clock and an uncorrelated relaxed clock with log-normal distribution (UCLN) [74]. In each case, we set the molecular clock rate to use a continuous time Markov chain rate reference prior [75]. For all six combinations of models, we performed path-sampling (PS) and stepping-stone sampling (SS) to estimate marginal likelihood [76, 77]. We sampled for 100 path steps with a chain length of 1 million, with power posteriors determined from evenly spaced quantiles of a Beta(alpha=0.3; 1.0) distribution. The Skyline tree prior provided a better fit than the two other (baseline) tree priors (Table D.2), so we used this tree prior for all further analyses. Using a constant or exponential tree prior, a relaxed clock provides a better model fit, as shown by the log Bayes factor when comparing the two clock models. Using a Skyline tree prior, the log Bayes factor comparing a strict and relaxed clock is smaller than it is using the other tree priors, and it is similar to the variability between estimated log marginal likelihood from PS and SS methods. We chose to use a relaxed clock for further analyses, but we also report key findings using a strict clock.

For the tree and tMRCA estimates in Figure 4.2, as well as the clock rate reported in main text, we ran BEAST with 400 million MCMC steps using the SRD06 substitution model, Skyline tree prior, and relaxed clock model. We extracted clock rate and tMRCA estimates, and their distributions, with Tracer v1.6.0 and identified the maximum clade credibility (MCC) tree using TreeAnnotator v1.8.4. We visualised the tree in FigTree v1.4.2 [68]. The reported credible intervals around estimates are 95% highest posterior density (HPD) intervals. When reporting substitution rate from a relaxed clock model, we give the mean rate (mean of the rates of each branch weighted by the time length of the branch). Additionally, for the tMRCA estimates in Figure 4.2C with a strict clock, we ran BEAST with the same specifications (also with 400M steps) except using a strict clock model. The resulting data are also used in the more comprehensive comparison shown in Figure D.3.

For the data with an outgroup in Figure D.3, we ran BEAST as specified above (with strict and relaxed clock models), except with 100 million steps and with outgroup sequences in the input alignment. The outgroup sequences were the same as those used to make the maximum likelihood tree. For the data excluding sample DOM_2016_MA-WGS16-020-SER in Figure D.3, we ran BEAST as specified above (with strict and relaxed clocks), except we removed the sequence of this sample from the input and ran 100 million steps.

We used BEAST v1.8.4 to estimate transition and transversion rates within the CDS and non-coding regions. The model was the same as above except that we used the Yang96 substitution model on the CDS, which uses GTR with gamma site heterogeneity and partitions codons into three partitions [78]; for the non-coding regions, we used a GTR substitution model with gamma

site heterogeneity and no codon partitioning. There were four partitions in total: one for each codon position and another for the non-coding region (5' and 3' UTRs combined). We ran this for 200 million steps. At each sampled step of the MCMC, we calculated substitution rates for each partition using the overall substitution rate, the relative substitution rate of the partition, the relative rates of substitutions in the partition, and base frequencies. In Figure D.4, we plot the means of these rates over the steps; the error bars shown are 95% HPD intervals of the rates over the steps.

We used BEAST v1.8.4 to reconstruct ancestral state at the root of the tree using CDS and non-coding regions. The model was the same as above except that, on the CDS, we used the HKY substitution model with gamma site heterogeneity and codons partitioned into three partitions (one per codon position). On the non-coding regions we used the same substitution model without codon partitioning. We ran this for 50 million steps and used TreeAnnotator v1.8.4 to find the state with the MCC tree. We selected the ancestral state corresponding to this state.

In all BEAST runs, we discarded the first 10% of states from each run as burn-in.

4.5.17 Principal Components Analysis

We carried out principal components analysis using the R package FactoMineR [79]. We imputed missing data with the package missMDA [80] and we show the results in Figure 4.2D.

4.5.18 Diagnostic Assay Assessment

We extracted primer and probe sequences from eight published RT-qPCR assays [26–31] and aligned them to our ZIKV genomes using Geneious version 9.1.7 [65]. We then tabulated matches

and mismatches to the diagnostic sequence for all outbreak genomes, allowing multiple bases to match where the diagnostic primer and/or probe sequence contained nucleotide ambiguity codes (Figure 4.3E).

4.5.19 Data Availability

Sequence data that support findings of this study have been deposited in NCBI GenBank [50] under BioProject accession PRJNA344504. Zika virus genomes have accession numbers KY014295–KY014327 and KY785409–KY785485. The dengue virus type 1 genome sequenced in this study has accession number KY829115.

4.6 Acknowledgements

We thank M. and L. Benioff for their vision and support; L. Brown, E. Lee, M. Giovanni, J. Levin-Allerhand and E. S. Lander for support and guidance; M. Schleicher, E. Lipscomb, A. Felix, A. Saltzman, and S. Donnelly for assistance with IRB and ethics processes; E. Mair, L. Nogelo and E. Carmean for legal counsel; T. Mason and the Broad Institute Genomics Platform for sequencing support; A. Matthews, S. Chapman, D. Neafsey, and B. Birren for management and guidance; O. Pybus and ZiBRA Project colleagues for sharing data before publication; D. Olson, E. Asturias, M. Salit, and E. Simon-Loriere for sharing samples and reagents; and E. Holmes, G. Bello, R. Tewhey, A. Piantadosi, C. Edwards and the Sabeti Laboratory for discussions and reading of the manuscript. We are indebted to Zika patients and clinical teams for making this work possible. Funding

was provided by: Marc and Lynne Benioff (P.C.S.); NIH NIAID U19AI110818 (Broad Institute); Howard Hughes Medical Institute (P.C.S.); Harvard University Burke Global Health Fellowship (P.C.S.); Broad Institute BroadNext10 program (A.G. and P.C.S.); AWS Cloud Credits for Research (P.C.S.); Conselho Nacional de Desenvolvimento Científico e Tecnológico (440909/2016-3) and Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (E-26/201.320/2016, E-26/201.332/2016, E-26/010.000194/2015) (P.T.B. and F.A.B.); NIH NIAID 1R01AI099210 (S.I. and S.F.M.); MIDAS-National Institute of General Medical Sciences U54GM111274 (M.E.H. and D.P.R.); NIH NIAID AI100190 (I.B. and L.G.); AEDES Network (I.B.) and Colombian Science, Technology and Innovation Fund of Sistema General de Regalías-BPIN 2013000100011 (L.V., R.M.G.R., M.C.M.M., and I.B.); ASTMH Shope Fellowship (K.G.B.); NSF DGE 1144152 (A.E.L.); PNPD/CAPES Postdoctoral Fellowship (E.D.); Fulbright-Colciencias Doctoral Scholarship (D.P.R.); NIH training grant 5T32AI007244-33 (N.D.G.); EU under grant agreements 278433-PREDEMICS and 643476-COMPARE (A.R.); and NIH NCATS CTSA UL1TR001114, NIH NIAID contract HHSN272201400048C, The Ray Thomas Foundation, and Pew Biomedical Scholarship (K.G.A.).

4.7 References

[1] Faria, N. R., Azevedo, R. d. S. d. S., Kraemer, M. U. G., et al. Zika virus in the Americas: Early epidemiological and genetic findings. *Science*, 352(6283):345–349, 2016.

[2] Pan American Health Organization. Zika: Epidemiological Report Honduras, 2017. URL http://www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=35137&Itemid=270.

[3] Centers for Disease Control and Prevention. First case of Zika virus reported in Puerto Rico, 2015. URL https://www.cdc.gov/media/releases/2015/s1231-zika.html.

[4] Pan American Health Organization. Zika: Epidemiological Report Dominican Republic, 2017. URL http://www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=35103&Itemid=270.

[5] World Health Organization. Zika situation report: Zika virus, Microcephaly and Guillain-Barré syndrome, 2017. URL http://apps.who.int/iris/bitstream/10665/254507/1/zikasitrep2Feb17-eng.pdf?ua=1.

[6] Reynolds, M. R., Jones, A. M., Petersen, E. E., et al. Vital Signs: Update on Zika Virus-Associated Birth Defects and Evaluation of All U.S. Infants with Congenital Zika Virus Exposure - U.S. Zika Pregnancy Registry, 2016. *MMWR. Morbidity and mortality weekly report*, 66(13):366–373, 2017.

[7] Dos Santos, T., Rodriguez, A., Almiron, M., et al. Zika Virus and the Guillain-Barré Syndrome - Case Series from Seven Countries. *New England Journal of Medicine*, 375(16):1598–1601, 2016.

[8] U.S. Food and Drug Administration. Zika Virus Response Updates from FDA, 2017. URL http://www.fda.gov/EmergencyPreparedness/Counterterrorism/MedicalCountermeasures/MCMIssues/ucm485199.htm#eua.

[9] Faria, N. R., Quick, J., Claro, I. M., et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*, 546(7658):406–410, 2017.

[10] Grubaugh, N. D., Ladner, J. T., Kraemer, M. U. G., et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature*, 546(7658):401–405, 2017.

[11] Metsky, H. C., Matranga, C. B., Wohl, S., et al. Zika virus evolution and spread in the Americas. *Nature*, 546(7658):411–415, 2017.

[12] de Vigilância em Saúde, S. Protocolo de vigilância e resposta à ocorrência de microcefalia, 2016. URL http://portalarquivos.saude.gov.br/images/pdf/2016/janeiro/22/microcefalia-protocolo-de-vigilancia-e-resposta-v1-3-22jan2016.pdf.

[13] Schieffelin, J. S., Shaffer, J. G., Goba, A., et al. Clinical illness and outcomes in patients with Ebola in Sierra Leone. *New England Journal of Medicine*, 371(22):2092–2100, 2014.

[14] Martina, B. E. E., Koraka, P., and Osterhaus, A. D. M. E. Dengue virus pathogenesis: an integrated view. *Clinical microbiology reviews*, 22(4):564–581, 2009.

[15] Mansuy, J. M., Mengelle, C., Pasquier, C., et al. Zika Virus Infection and Prolonged Viremia in Whole-Blood Specimens. *Emerging Infectious Diseases*, 23(5):863–865, 2017.

[16] Fourcade, C., Mansuy, J. M., Dutertre, M., et al. Viral load kinetics of Zika virus in plasma, urine and saliva in a couple returning from Martinique, French West Indies. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology*, 82:1–4, 2016.

[17] Paz-Bailey, G., Rosenberg, E. S., Doyle, K., et al. Persistence of Zika Virus in Body Fluids — Preliminary Report. *New England Journal of Medicine*, 2017.

[18] Nayak, S., Lei, J., Pekosz, A., Klein, S., and Burd, I. Pathogenesis and Molecular Mechanisms of Zika Virus. *Seminars in Reproductive Medicine*, 34(05):266–272, 2016.

[19] Sardi, S. I., Somasekar, S., Naccache, S. N., et al. Coinfections of Zika and Chikungunya Viruses in Bahia, Brazil, Identified by Metagenomic Next-Generation Sequencing. *Journal of clinical microbiology*, 54(9):2348–2353, 2016.

[20] Fauci, A. S. and Morens, D. M. Zika Virus in the Americas — Yet Another Arbovirus Threat. *New England Journal of Medicine*, 374(7):601–604, 2016.

[21] Quick, J., Grubaugh, N. D., Pullan, S. T., et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols*, 12(6):1261–1276, 2017.

[22] Matranga, C. B., Andersen, K. G., Winnicki, S., et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol*, 15(519):519, 2014.

[23] Sall, A. A., Faye, O., Diallo, M., et al. Yellow fever virus exhibits slower evolutionary dynamics than dengue virus. *J. Virol.*, 84(2):765–772, 2010.

[24] Pan American Health Organization. Epidemiological Update: Zika virus infection, 2015. URL http://www.paho.org/hq/index.php?option=com_docman&task=doc_view&Itemid=270&gid=32021&lang=en.

[25] Nunes, M. R. T., Faria, N. R., de Vasconcelos, J. M., et al. Emergence and potential for spread of Chikungunya virus in Brazil. *BMC medicine*, 13:102, 2015.

[26] Pyke, A. T., Daly, M. T., Cameron, J. N., et al. Imported zika virus infection from the cook islands into australia, 2014. *PLoS Currents*, 6, 2014.

[27] Lanciotti, R. S., Kosoy, O. L., Laven, J. J., et al. Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. *Emerging Infectious Diseases*, 14(8): 1232–1239, 2008.

[28] Faye, O., Faye, O., Dupressoir, A., et al. One-step RT-PCR for detection of Zika virus. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology*, 43 (1):96–101, 2008.

[29] Faye, O., Faye, O., Diallo, D., et al. Quantitative real-time PCR detection of Zika virus and evaluation with field-caught mosquitoes. *Virol. J.*, 10:311, 2013.

[30] Balm, M. N. D., Lee, C. K., Lee, H. K., et al. A diagnostic polymerase chain reaction assay for Zika virus. *Journal of Medical Virology*, 84(9):1501–1505, 2012.

[31] Tappe, D., Rissland, J., Gabriel, M., et al. First case of laboratory-confirmed Zika virus infection imported into Europe, November 2013. *Eurosurveillance*, 19(4), 2014.

[32] Tsetsarkin, K. A., Vanlandingham, D. L., McGee, C. E., and Higgs, S. A single mutation in chikungunya virus affects vector specificity and epidemic potential. *PLoS pathogens*, 3(12): e201, 2007.

[33] Piantadosi, A., Chohan, B., Panteleeff, D., et al. HIV-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response. *AIDS*, 23 (5):579–587, 2009.

[34] Villabona-Arenas, C. J., Mondini, A., Bosch, I., et al. Dengue Virus Type 3 Adaptive Changes during Epidemics in São Jose de Rio Preto, Brazil, 2006–2007. *PLoS ONE*, 8(5):e63496, 2013.

[35] Brinton, M. A. and Basu, M. Functions of the 3' and 5' genome RNA regions of members of the genus Flavivirus. *Virus research*, 206:108–119, 2015.

[36] Duchêne, S., Ho, S. Y. W., and Holmes, E. C. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC evolutionary biology*, 15:36, 2015.

[37] Corman, V. M., Rasche, A., Baronti, C., et al. Clinical comparison, standardization and optimization of Zika virus molecular detection. *Bull. World Health Organ.*, 2016.

[38] Nowak, M. A. What is a quasispecies? *Trends in Ecology & Evolution*, 7(4):118–121, 1992.

[39] Domingo, E. and Holland, J. J. RNA VIRUS MUTATIONS AND FITNESS FOR SURVIVAL. *Annual Review of Microbiology*, 51(1):151–178, 1997.

[40] Duffy, S., Shackelton, L. A., and Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature*, 9(4):267–276, 2008.

[41] Posada-Cespedes, S., Seifert, D., and Beerenwinkel, N. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus research*, 239:17–32, 2016.

[42] Nowak, M., Anderson, R., McLean, A., et al. Antigenic diversity thresholds and the development of AIDS. *Science*, 254(5034):963–969, 1991.

[43] Gaschen, B., Taylor, J., Yusim, K., et al. Diversity Considerations in HIV-1 Vaccine Selection. *Science*, 296(5577):2354–2360, 2002.

[44] Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., and Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions within a viral population. *Nature*, 439(7074):344, 2006.

[45] Gire, S. K., Goba, A., Andersen, K. G., et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[46] Park, D. J., Dudas, G., Wohl, S., et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526, 2015.

[47] Emmett, K. J., Lee, A., Khiabanian, H., and Rabadan, R. High-resolution Genomic Surveillance of 2014 Ebolavirus Using Shared Subclonal Variants. *PLoS Currents*, 7, 2015.

[48] Morlan, J. D., Qu, K., and Sinicropi, D. V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS ONE*, 7(8):e42882, 2012.

[49] Worobey, M., Watts, T. D., McKay, R. A., et al. 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature*, 539(7627):98–101, 2016.

[50] Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. GenBank. *Nucleic acids research*, 44(D1):D67–72, 2016.

[51] Tomkins-Tinch, C., Ye, S., Metsky, H., et al. viral-ngs, 2016. URL doi:10.5281/zenodo.200428.

[52] Wood, D. E. and Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

[53] O'Leary, N. A., Wright, M. W., Brister, J. R., et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–45, 2016.

[54] Aurrecoechea, C., Brestelli, J., Brunk, B. P., et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic acids research*, 37(Database issue):D539–43, 2009.

[55] Yarza, P., Richter, M., Peplies, J., et al. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.*, 31(4):241–250, 2008.

[56] Brister, J. R., Ako-Adjei, D., Bào, Y., and Blinkova, O. NCBI viral genomes resource. *Nucleic acids research*, 43(Database issue):D571–7, 2015.

[57] Grabherr, M. G., Haas, B. J., Yassour, M., et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[58] Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[59] Hyatt, D., LoCascio, P. F., Hauser, L. J., and Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, 28(17):2223–2230, 2012.

[60] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 44(D1):D7–19, 2016.

[61] R Core Team. R: A Language and Environment for Statistical Computing, 2016. URL https://www.R-project.org/.

[62] Cribari-Neto, F. and Zeileis, A. Beta Regression in R. *J. Stat. Softw.*, 34(1):1–24, 2010.

[63] Katoh, K. and Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, 2013.

[64] Li, H., Handsaker, B., Wysoker, A., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[65] Kearse, M., Moir, R., Wilson, A., et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12): 1647–1649, 2012.

[66] Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *PNAS*, 89(22):10915–10919, 1992.

[67] Guindon, S., Dufayard, J.-F., Lefort, V., et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3):307–321, 2010.

[68] Rambaut, A. FigTree, 2014. URL http://tree.bio.ed.ac.uk/software/figtree/.

[69] Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*, 2(1):vew007, 2016.

[70] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*, 29(8):1969–1973, 2012.

[71] Shapiro, B., Ho, S. Y. W., Drummond, A. J., et al. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol Biol Evol*, 28(2):879–887, 2011.

[72] Shapiro, B., Rambaut, A., and Drummond, A. J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol*, 23(1):7–9, 2006.

[73] Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*, 22(5):1185–1192, 2005.

[74] Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, 4(5):e88, 2006.

[75] Ferreira, M. A. R. and Suchard, M. A. Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.*, 36(3):355–368, 2008.

[76] Baele, G., Lemey, P., Bedford, T., et al. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol*, 29(9): 2157–2167, 2012.

[77] Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A., and Lemey, P. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol Biol Evol*, 30(2):239–243, 2013.

[78] Yang, Z. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J Mol Evol*, 42(5):587–596, 1996.

[79] Lê, S., Josse, J., and Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.*, 2008.

[80] Josse, J. and Husson, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *J. Stat. Softw.*, 70(1):1–31, 2016.

Chapter 5

Genomic Epidemiology Reveals Ongoing Mumps Transmission in the United States

Preface

This final chapter covers an ongoing exploration of the recent mumps virus (MuV) outbreak in the United States. Following our work on reconstructing transmission during the EBOV outbreak in Nigeria, we hoped to apply a similar approach to another outbreak for which we could obtain comprehensive contact tracing data. The close proximity of the MuV outbreak — hundreds of cases were reported in Massachusetts in 2016–2017 — and the efforts of the local university health services and Massachusetts Department of Public Health (MDPH) to trace patients and record epidemiological information made it a perfect subject for this type of analysis. Proximity to the outbreak also eliminated issues related to sample transfer, ensuring the availability of high-quality samples for sequencing. Additionally, access to high-quality samples created an opportunity to further study within-host variation, which we could do only in a limited fashion with the lower quality

EBOV samples from Nigeria.

Beyond these methodological goals, we hoped to use MuV sequences to answer public health questions about the nature of the MuV outbreak. We sought to understand the presence of MuV in highly vaccinated university communities, how the virus got into Massachusetts, and the extent to which genomic data could be used to trace transmission within and between university communities. These questions are answered in the text below, which has been adapted from a draft manuscript we plan to submit for publication in the coming weeks.

While I have been the main driver of this mumps project, there are many people who have contributed immensely to the work presented below. I optimized our sequencing protocol (see [1]) for MuV with Anne Piantadosi, Katie Siddle, and Chris Matranga, and sequenced over 200 MuV samples with help from Katie Siddle, Bettina Bankamp, Rickey Shah, and James Qu. Bridget Chak and I worked closely with Meagan Burns (MDPH) to collate epidemiological data from the outbreak for use in transmission analyses. Hayden Metsky performed most of the phylogenetic analysis in BEAST, as well as a meta-analysis of SH gene data. Steve Schaffner and I analyzed within- and between-host variation of MuV, and Anne Piantadosi and Bridget Chak have been instrumental in exploring the potential of these mutations to confer vaccine escape. This project would not have happened without our close collaboration with the MDPH, fostered by my advisor Pardis Sabeti, Nathan Yozwiak, Danny Park, and others, as well as the vision of Sandy Smole, Larry Madoff and our other MDPH collaborators. Yonatan Grad has also been extremely important to both envisioning the project at its start, as well as discussing every aspect of the analysis. The manuscript below has been primarily written by me, Bronwyn MacInnis, Steve Schaffner, Nathan Yozwiak, Hay-

den Metsky, and Anne Piantadosi and, in our opinion, demonstrates the true potential of genomics to influence public health measures during and after viral outbreaks.

5.1 Abstract

Despite widespread vaccination, thousands of mumps cases were reported in the United States in 2016–2017, including hundreds in Massachusetts, primarily in college settings. We generated 203 whole genome sequences of mumps virus (MuV) from Massachusetts and 15 other states to understand the dynamics of mumps spread locally and nationally, as well as to search for mutations associated with vaccine escape. We observed multiple lineages of MuV circulating within Massachusetts during the outbreak, evidence for multiple introductions of the virus to the state, and extensive geographic movement of MuV within the country on short time scales. We found no association between MuV lineage and vaccine status, and little evidence that mutations that arose during this outbreak contributed to vaccine escape. Combining epidemiological and genomic data, we observed multiple co-circulating clades within individual universities as well as spillover into the local community. We also used publicly available sequences from a single gene to estimate migration between world regions and to place this outbreak in a global context, but found this short sequence to be inadequate for tracing detailed transmission. Our findings suggest continuous, often undetected circulation of mumps both locally and nationally, and highlight the value of combining genomic and epidemiological data to track viral disease transmission at high resolution.

5.2 Introduction

An unusually large number of mumps cases were reported in the United States in 2016 and 2017, despite high rates of vaccination [2, 3]. Mumps incidence declined by more than 98% after introduction of the mumps vaccine in 1967. Case counts briefly rose again in the mid-1980s, but continued to drop after a second dose of the vaccine was recommended in 1989, and in the early 2000s only a few hundred cases were observed annually in the United States [2]. Low nationwide incidence was interrupted by a large outbreak (>5,000 cases) in the midwestern United States in 2006 [4], followed by another period of low incidence with minor outbreaks until 2016.

Massachusetts was one of the states with a large mumps outbreak in 2016–2017. Over 250 cases were reported to the MDPH in 2016 and more than 170 in 2017, far exceeding the usual state incidence of <10 cases per year [5] (Figure 5.1A–B). As seen in other recent outbreaks, most cases were associated with universities and other close contact settings [4, 6]. Most cases were in college-aged individuals and at least 65% of cases were in individuals with the recommended two doses of the Measles-Mumps-Rubella (MMR) vaccine (Table E.3).

5.3 Results

We used genomic epidemiology to investigate the spread of MuV in Massachusetts and on national and international scales to better understand routes of mumps transmission, as well as to determine whether mutations associated with vaccine escape could be identified. We generated 203 whole genomes (160 from Massachusetts and 43 from 15 other states) using a combination of unbi-

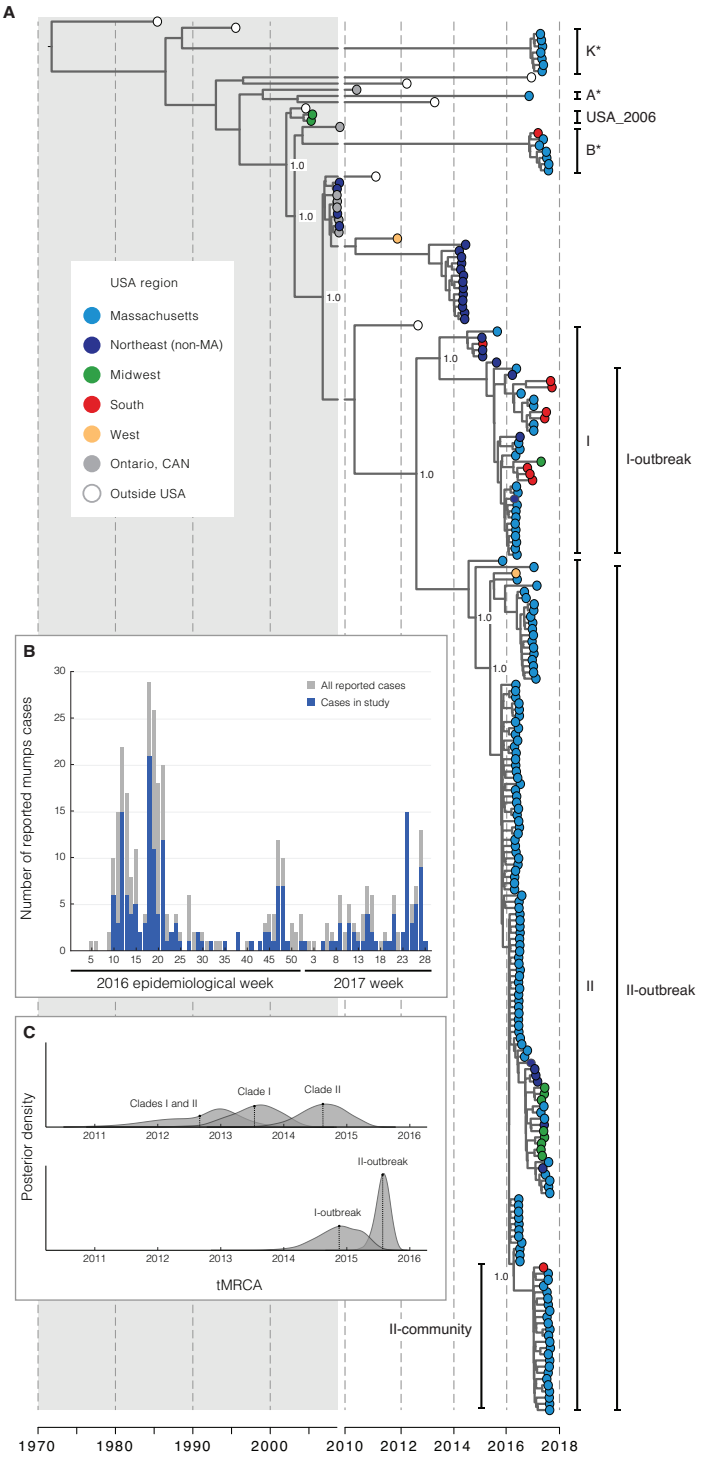
ased and capture-based sequencing approaches [1, 7] (see Methods, Section 5.5, and Table E.1). The genomes had a median of 99.48% unambiguous base calls and a median depth of 176× (Figure E.1). We also sequenced a set of 29 PCR-negative samples from patients with suspected mumps to determine if we could detect MuV or other viruses that may explain their symptoms. We saw evidence for MuV in one sample and identified four other viruses known to cause upper respiratory symptoms (at least two of which are known to cause parotitis) in four separate samples [8–10] (Table E.2).

5.3.1 Multiple Co-Circulating Mumps Virus Lineages

To understand how the Massachusetts outbreak fit into the larger context of mumps in the United States, we performed a phylogenetic analysis on our dataset together with all genotype G whole genomes available from NCBI GenBank [11] (n=25). The resulting phylogeny (Figure 5.1A) suggests that most mumps viruses from the recent Massachusetts outbreak descend from those in the 2006 U.S. outbreak: their genomes fall within the same clade as the 2006 samples, and the root of the clade lies within or very close to the 2006 samples. This conclusion is supported by a root-to-tip analysis (Figure E.2B–C). Within this clade, we also find samples from the United States in intervening years (2009–2015), suggesting sustained transmission of this clade within the United States. We also see that mumps viruses from Massachusetts are interspersed with sequences from across the United States (Figure 5.1A), providing evidence for extensive geographic movement of MuV within the country on short timescales. A principal components analysis of MuV sequences (Figure E.2D) similarly shows that sequences from other regions cluster with ones from Massachusetts.

Figure 5.1: Mumps cases in Massachusetts and circulation of MuV in the United States. (**A**) Maximum clade credibility tree of 225 genotype G whole genome MuV sequences, including 200 generated in this study. Labels on selected internal nodes indicate posterior support. Relevant clades identified here are labeled. Clades I and II contain 91% of the samples from the 2016–2017 Massachusetts outbreak. I-outbreak and II-outbreak are the smallest clades within I and II, respectively, that contain all samples from the 2016–2017 Massachusetts outbreak. K* contains samples associated with Institution K other than those in clades I and II; the same is true for B* and A*. II-community is a clade comprised primarily of samples associated with a geographic community. (**B**) Number of reported mumps cases by epidemiological week in Massachusetts (gray) and in this study (blue); samples included in this study are representative of all cases (see Table E.3). (**C**) Probability distributions for the tMRCA of selected sequences labeled in (A) (see Table E.4 for additional clades). Dotted lines indicate the mean of each distribution.

Figure 5.1: Continued



We see clear evidence for several introductions of MuV into Massachusetts, including two distinct MuV lineages (clades I and II) descended from the 2006 outbreak that diverged late in 2012 (Figure 5.1C, Table E.4). To estimate when the two primary clades entered Massachusetts, we calculated the time to the most recent common ancestor (tMRCA) of each using only samples from the 2016–2017 Massachusetts outbreak (I-outbreak tMRCA = July 2015, II-outbreak tMRCA = November 2014) (Figure 5.1C, Table E.4). While most MuV in Massachusetts falls within this 2006 lineage, we are able to detect small transmission clusters within the outbreak that resulted from importation events from elsewhere: three clades contained MuV sequences distinct from the majority of Massachusetts sequences, and each had at least one MuV sequence from a patient with foreign travel history during the incubation period of the virus.

The sequence data resolved details of the outbreak at finer scales as well, showing that multiple clades were also co-circulating within individual academic institutions. For example, there were two viral lineages in samples associated with Institution K (clades II and K*) and multiple viral lineages in samples from Institution B–associated individuals (clades I, II, and B*) (Figure E.3A). Thus, what appeared to be a single outbreak across multiple institutions was shown by sequence data to be multiple overlapping outbreaks.

5.3.2 Mumps Virus Transmission within Massachusetts

Sequence data also allowed us to identify a spillover event from one institution into the larger community. Before genomic data was available, cases associated with Institution A and with a geographic community (clade II-community) were inferred to be separate outbreaks due to the dif-

ferent populations affected (mostly students versus adults with no reported university connection) and an apparent five month gap between the two sets of cases. From the phylogeny, however, it is clear that these two groups of cases are related and that the community-associated cases represent a spillover from Institution A into the broader population (Figure 5.2A). Additional epidemiological investigation revealed three infected individuals associated with both communities who could have served as transmission links. The long gap between these two sets of linked cases suggests local undetected mumps circulation and, in line with the picture seen above, sustained transmission in the United States that is only sporadically detected.

The comprehensive sampling at Institution A again allowed us to combine genomic and epidemiology data to better understand the outbreak dynamics, this time by estimating the number of mumps introductions into the university. Detailed epidemiological data alone were inadequate to determine whether the initial effective reproductive number, $R_E(t{=}0)$, within the university exceeded the critical threshold of 1.0 (Figures 5.2B left, E.4). However, incorporating the number of distinct viral lineages observed markedly improved our ability to infer transmission dynamics within the institution, supporting an estimate of five (95% CI: 4–18) distinct introductions that were each expected to cause $R_E(t{=}0){=}1.70$ (95% CI: 1.50–1.91) secondary cases (Figure 5.2B right). That $R_E(t{=}0)$ is well above one has implications for the required reach of reactive vaccination campaigns in at-risk populations: in this case, vaccination would need to reach 59% (52–67%) to effectively curtail transmission.

We next investigated the usefulness of sequence data to supplement epidemiological data at the finest analysis scale: reconstructing individual transmission chains. To determine the extent

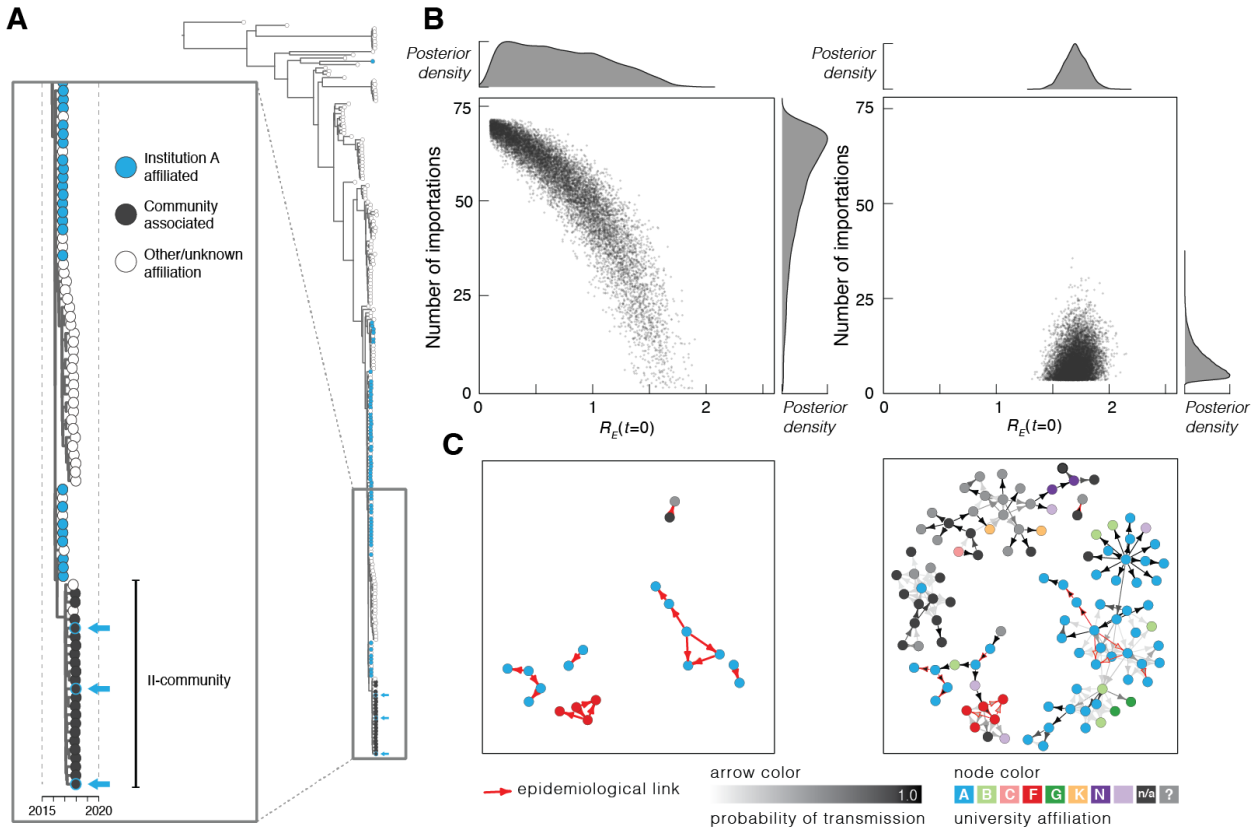


Figure 5.2: Epidemiological modeling and transmission reconstruction. (**A**) Zoom view of the clade II-community and its ancestors in the maximum clade credibility tree from Figure 5.1A, colored by academic institution affiliation. Arrows highlight samples from three individuals affiliated with both the II-community and Institution A. (**B**) Number of importations into Institution A calculated without (left) and with (right) viral genetic information as input. Each point represents a sample from the posterior distribution of $R_E(t{=}0)$ and the number of introductions, based on simulated transmission dynamics under varying values of these parameters. (**C**) Transmission reconstruction of individuals within clade II-outbreak; samples are colored by institution affiliation. Left: reconstruction using epidemiological data only; all individuals in clade II-outbreak with known epidemiological links (red arrows) are shown. Right: reconstruction using MuV genomes and collection dates. Arrow shading indicates probability of direct transmission between individuals, and individuals with no estimated links to other samples in this clade are not shown. Arrows outlined in red represent transmission events identified by both genomic and epidemiological data.

to which genetic data can be used to infer epidemiological links, we examined the genetic distance between samples with a known epidemiological link (i.e., samples likely part of the same transmission chain). All samples with a known link were genetically similar (Figure E.5A), and genetic distance was a good predictor of epidemiological linkage (Figure E.5B). Given this, we used genetic data (along with sampling dates) to reconstruct transmission chains during the outbreak (Figure 5.2C), focusing on samples within clade II-outbreak. This analysis correctly inferred all known epidemiological links.

Within-host variants (intrahost variants, or iSNVs) shared between samples can provide additional information about transmission chains during viral outbreaks [12–14]. In our dataset, we identified iSNVs in 52% of samples. Most of the samples without iSNVs had low sequencing coverage, though we did observe this lack of iSNVs in samples with high sequencing coverage as well. Ultimately, however, iSNVs proved uninformative in understanding MuV transmission. We did not find evidence for shared iSNVs in the five pairs of known direct contacts or above 0.1% frequency between any samples. These data suggest that the MuV transmission bottleneck may be small enough to preclude shared within-host variation. We note, however, that in two patients for whom we had multiple samples (in both cases collected nine days apart), the MuV genomes from the same patient differed at one nucleotide position. In one pair, different alleles were fixed in the two genomes; in the other, the genome from the second time point had an iSNV at one position (56% alternate allele, 44% matching the other genome).

5.3.3 Variants Potentially Associated with Vaccine Escape

We also looked for variation between genomes in our dataset that could be associated with vaccine or immune evasion. Site-specific d_N/d_S analysis of all available genotype G MuV genomes provided no strong evidence for positive selection at any site (Figure 5.3A); overall, the greatest selective constraint was observed in the polymerase (L) and structural proteins (Figure 5.3B). No fixed nucleotide substitutions were associated with vaccine status or time since vaccination (Figures 5.3C, E.3B).

We also investigated changes in immunogenic regions of the mumps genome, particularly the hemagglutinin (HN) protein, which is the primary target of neutralizing antibodies (NAb). We identified 32 fixed amino acid substitutions in HN between the strains in our data set and the Jeryl Lynn vaccine strain (the vaccine strain used in the United States). These included substitutions within one putative and two experimentally-defined NAb epitopes and both a gain and a loss of potential N-linked glycosylation site (Figure 5.3D) [15–19]. We also observed eight fixed amino acid substitutions between our sequences and the Jeryl Lynn strain in the hypervariable hydrophilic C-terminus of the nucleoprotein (NP), which has also been demonstrated to contain NAb targets [20]. At all but two of these sites in HN and NP, the allele in our sequences is the same as in a strain isolated in Iowa in 2006 (GenBank accession: JX287385.1); MuV from this outbreak has been shown to be neutralized by sera from both vaccinated and naturally-infected individuals [21]. Our strains differed from the Iowa 2006 strain at HN positions 336 (in a known NAb epitope) and 474 (in a putative NAb epitope). At both of these positions, most genotype G sequences share the same allele

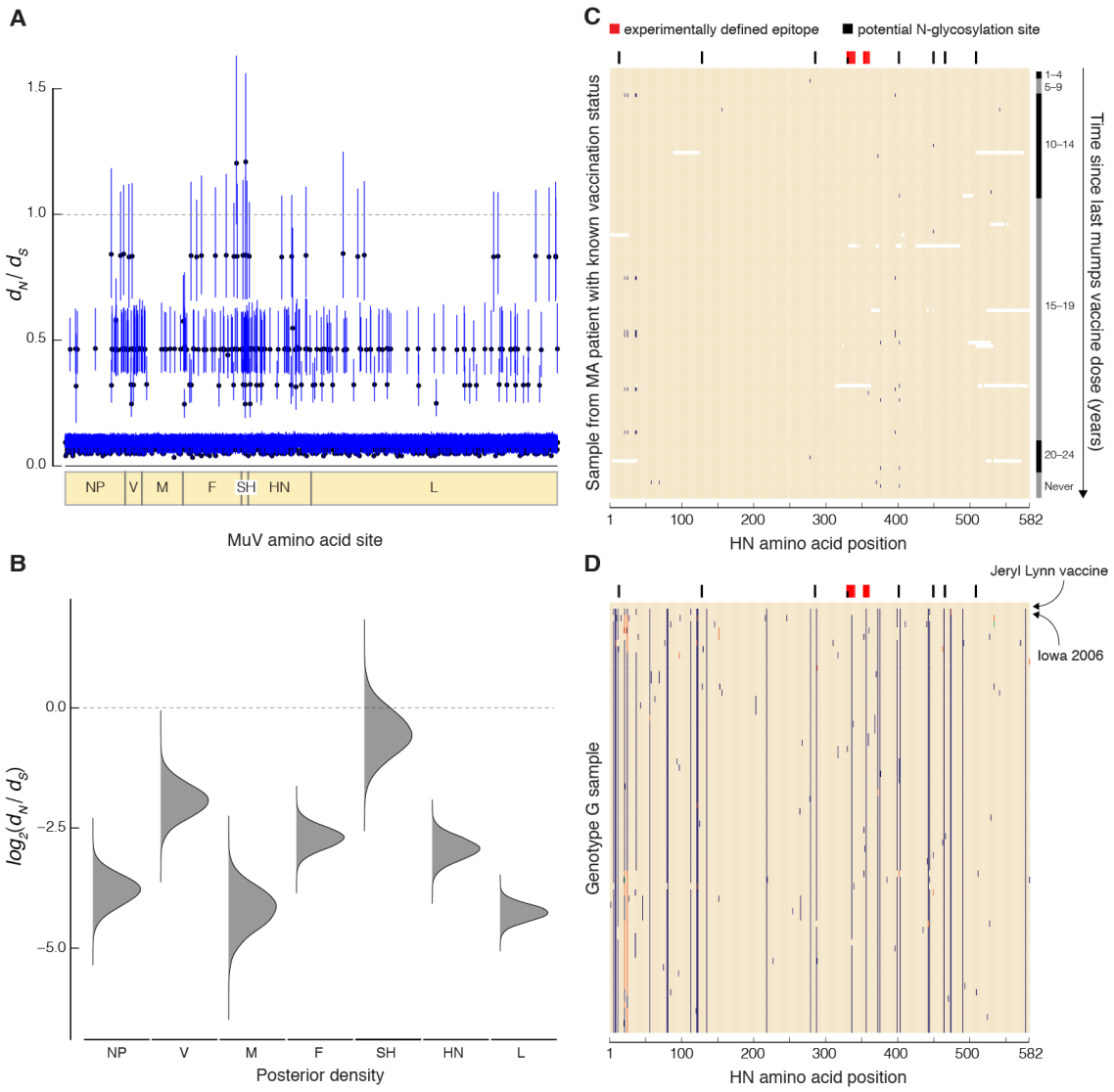


Figure 5.3: Amino acid substitution in the MuV genome. (**A**) Estimate of d_N/d_S at each amino acid site in MuV coding regions, calculated across all 225 genotype G genomes used in this study. (**B**) Posterior density of d_N/d_S in each MuV gene, using the same dataset. (**C**) Variation in genomes generated in this study. Each row represents one of the 119 MuV HN amino acid sequences from the individuals in our study who had known vaccination status. Samples are displayed in order of descending time since last MMR vaccine dose. Colored variants indicate variation from the consensus of all included sequences. (**D**) Variation in all published genotype G HN sequences. Each row represents one of the 456 publicly available MuV genotype G HN sequences (including from genomes generated in this study). Identical sequences are collapsed and then grouped by hierarchical clustering. In both panels, amino acid substitutions relative to the Jeryl Lynn vaccine strain are highlighted in dark blue, with orange indicating a second variant allele and green indicating a third. Red bars indicate experimentally-identified neutralizing antibody epitopes, and black bars indicate potential N-glycosylation sites.

as our sequences (Figure 5.3D), indicating that the Iowa 2006 strain is an outlier at these sites, and suggesting the need for further studies testing the neutralization susceptibility of strains containing the common alleles.

5.3.4 Global Spread of Mumps Virus

Finally, we attempted to understand MuV circulation in its global context; for this purpose, we analyzed the 316-nucleotide small hydrophobic (SH) gene, historically the target of most studies of MuV and the basis for defining MuV genotypes [22, 23]. The dataset consisted of 3,646 sequences available on NCBI GenBank [11] from around the world, including those from genomes generated in this study (Figures 5.4A, E.6). We calculated tMRCA for each of the 11 included genotypes and found that genotype A, the strain used in most vaccines (including the Jeryl Lynn vaccine used in the United States), coalesces the earliest (Figure 5.4B). This genotype appears to have stopped circulating (previously noted in Jin et al. [24]), raising the possibility that vaccination contributed to the disappearance of that genotype.

We performed a phylogeographic analysis on these SH sequences to understand movement of MuV between world regions. To reduce temporal and geographic sampling biases, we looked at samples collected since 2010 in four well-sampled regions (United States, Europe, East Asia, and South/Southeast Asia). We see significant MuV migration between the United States and Europe (Figures 5.4C, E.6C) and find that most introductions to the United States are from Europe, although there is also support for mumps jumps from East Asia and South/Southeast Asia to the United States. Likewise, most introductions to Europe appear to be from the United States. We

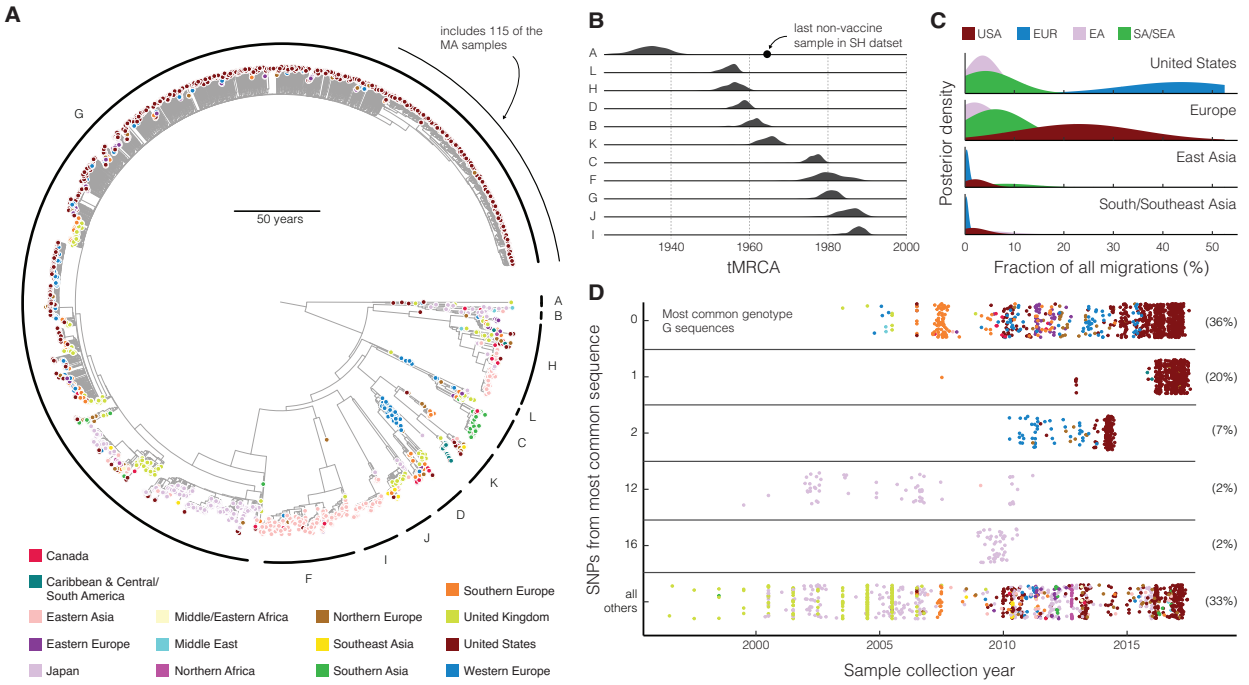


Figure 5.4: Global spread of MuV using SH gene sequences. (**A**) Maximum clade credibility tree of 3,646 publicly available SH gene sequences, including 193 generated in this study. Sample tips are colored by one of 15 world regions. (**B**) Probability distributions for the tMRCA of each of the 11 included genotypes. The date of the most recent genotype A clinical sample is indicated, excluding samples likely to contain a MuV vaccine. (**C**) Migration between four global regions. In each plot, the migration destination is indicated. Each panel shows a posterior probability density, taken across resampled input, of the fraction of all reconstructed migrations that occur to the destination from each of the other three sources (see Methods, Section 5.5, for details regarding geographic and temporal resampling of sequences). (**D**) Identical genotype G sequences over time. Each dot represents a sample and each row contains samples with identical SH sequences, except the bottom, which includes samples whose sequences are distinct from those in the above five categories. Numbers on the right of each row are the percentage of all genotype G samples found in that row.

also found that recent sequences from the United States have primarily European ancestry, and vice versa (Figure E.6D). Notably, recent sequences from East Asia and South/Southeast Asia have minimal external ancestry, indicating relatively little spread to these regions (Figure 5.4C).

It is important to note that the SH gene is less than 3% of the MuV genome and represents only a fraction of MuV genetic diversity. While SH sequences are useful for analyzing diversity on a global scale, their lack of variability (Figure 5.4D) may inflate estimates of migration. Moreover, lack of variability makes it difficult or impossible to reconstruct MuV spread on smaller temporal and geographic scales (e.g., analysis using SH sequences alone would not have shown a genetic link between Institution A and the geography community, Figure E.7), highlighting the importance of whole genome sequencing.

5.4 Discussion

Here we show the importance of pairing genomics with epidemiology to understand the spread of a mumps outbreak. Despite detailed epidemiological data collected in the outbreak, MuV sequence data were required to identify multiple co-circulating strains within Massachusetts (and even within single institutions), to connect two outbreaks previously thought to be separate, and to accurately estimate R_E within one institution. Additionally, while SH gene sequencing reveals global trends in MuV circulation, analysis based on a single, short gene would not be adequate for producing this kind of detailed picture of how mumps is spreading in the United States (Figure E.7). Not only do these findings reveal transmission patterns in this particular outbreak, they also

suggest the presence of undetected mumps circulating within Massachusetts and the United States, and a possible high asymptomatic burden despite widespread immunization. This has implications for vaccination use and recommendations, already under consideration by the U.S. Centers for Disease Control and Prevention [25].

Although mumps is not deadly, it does serve as a good model for the use of genomic data in a more severe viral outbreak, in which reconstructing transmission and analyzing spread may have great public health significance. We expect that applying detailed genomic and epidemiological data to viral outbreaks will play an increasingly important role in surveillance and response.

5.5 Methods

5.5.1 Data Availability

The 203 MuV whole genome sequences generated in this study, as well as nine low quality sequences not included in the analysis, are available on NCBI GenBank under BioProject accession PRJNA394142 (accession numbers MF965196–MF965318 and MG986380–MG986468).

5.5.2 Ethics Statement and Study Subjects

The study protocol was approved by the MDPH, Centers for Disease Control and Prevention (CDC), and Massachusetts Institute of Technology (MIT) Institutional Review Boards (IRBs). Harvard University Faculty of Arts and Sciences and the Broad Institute ceded review of sequencing and secondary analysis to the MDPH IRB through authorization agreements. The MDPH IRB waived

informed consent given this research met the requirements pursuant to 45 C.F.R. 46.116 (d).

Buccal swab samples were obtained from suspected mumps cases at the MDPH and CDC for testing. Both cohorts included only leftover clinical diagnostic samples.

5.5.3 Viral RNA Isolation

Sample inactivation and RNA extraction were performed at the MDPH, Broad Institute, and CDC. At MDPH, viral samples were inactivated by adding 300 μL Lysis/Binding Buffer (Roche) to 200 μL sample, vortexing for 15 seconds, and incubating lysate at room temperature for 30 minutes. RNA was then extracted following the standard external lysis extraction protocol from the MagNA Pure LC Total Nucleic Acid Isolation Kit (Roche) using a final elution volume of 60 μL. At the Broad Institute, samples were inactivated by adding 252 μL Lysis/Binding Buffer (ThermoFisher) to 100 μL sample and RNA was then extracted following the standard protocol from the MagMAX Pathogen RNA/DNA Kit (ThermoFisher) using a final elution volume of 75 μL. At the CDC, RNA extraction followed the standard protocol from the QiaAmp Vira RNA mini kit (Qiagen).

5.5.4 PCR Diagnostic Assays Performed at MDPH and CDC

Diagnostic tests for presence of MuV were performed at MDPH and CDC using the CDC Real-Time (TaqMan) RT-PCR Assay for the Detection of Mumps Virus RNA in Clinical Samples [26]. Each sample was run in triplicate using both the Mumps N Gene assay (MuN) and RNase P (RP) assays using the standard CDC protocol and primers. RT-PCR was performed on the Applied Biosystems 7500 Fast Real-Time PCR system or Applied Biosystems Prism 7900HT Sequence Detection System

instrument.

5.5.5 PCR Quantification Assays Performed at Broad Institute

MuV RNA was quantified at the Broad Institute using the Power SYBR Green RNA-to-Ct 1-Step qRT-PCR assay (Life Technologies) and CDC MuN primers. The 10 μL assay mix included 3 μL RNA, 0.3 μL each MuV forward and reverse primers at 5 μM concentration, 5 μL 2x Power SYBR RT-PCR Mix, and 0.08 μL 125$\times$ RT Enzyme Mix. The cycling conditions were 48°C for 30 min and 95°C for 10 min, followed by 45 cycles of 95°C for 15 sec and 60°C for 30 sec with a melt curve of 95°C for 15 sec, 55°C for 15 sec, and 95°C for 15 sec. RT-PCR was performed on the ThermoFischer QuantStudio 6 instrument. To determine viral copy number, we used a double-stranded gene fragment (IDT gBlock) as a standard. This standard is a 171 bp fragment of the MuV genome (GenBank accession: NC_002200) including the amplicon (sequence: GGA TCG ATG CTA CAG TGT ACT AAT CCA GGC TTG GGT GAT GGT CTG TAA ATG TAT GAC AGC GTA CGA CCA ACC TGC TGG ATC TGC TGA TCG GCG ATT TGC GAA ATA CCA GCA GCA AGG TCG CCT GGA AGC AAG ATA CAT GCT GCA GCC AGA AGC CCA AAG GTT GAT TCA AAC).

23S rRNA content in samples was quantified using the same Power SYBR Green RNA-to-Ct 1-Step qRT-PCR assay kit and cycling conditions. Primers were used to amplify a 183 bp universally conserved region of the 23S rRNA (fwd: 93a - GGG TTC AGA ACG TCG TGA GA, rev: 97ar - CCC GCT TAG ATG CTT TCA GC) [27]. To determine viral copy number, we used a double-stranded gene fragment (IDT gBlock) as a standard. This standard is a 214 bp fragment of the Streptococcus HTS2 genome (GenBank accession: NZ_CP016953) (sequence: AGC GGC ACG CGA GCT GGG TTC

AGA ACG TCG TGA GAC AGT TCG GTC CCT ATC CGT CGC GGG CGT AGG AAA TTT GAG AGG ATC TGC TCC TAG TAC GAG AGG ACC AGA GTG GAC TTA CCG CTG GTG TAC CAG TTG TCT CGC CAG AGG CAT CGC TGG GTA GCT ATG TAG GGA AGG GAT AAA CGC TGA AAG CAT CTA AGT GTG AAA CCC ACC TCA AGA T).

5.5.6 Bacterial rRNA Depletion

Bacterial rRNA was depleted from some RNA samples using the Ribo-Zero Bacteria Kit (Illumina). At the hybridization step, the 40 μL reaction mix included 5 μL RNA sample, 4 μL Ribo-Zero Reaction Buffer, 8 μL Ribo-Zero Removal Solution, 22.5 μL water, and 0.5 μL synthetic RNA (25 fg) used to track potential cross-contamination (ERCC, gift from M. Salit, NIST). Bacterial rRNA-depleted samples were purified using 1.8$\times$ volumes Agencourt RNAClean XP beads (Beckman Coulter) and eluted in 10 μL water for cDNA synthesis.

5.5.7 Illumina Library Construction and Sequencing

cDNA synthesis was performed as described in previously published RNA-seq methods [1]. In samples where bacterial rRNA was not depleted, 25 fg synthetic RNA was added at the beginning of cDNA synthesis to track sample cross-contamination. Positive control libraries were prepared from a mock MuV sample in which a 1:100 dilution of Enders strain MuV was spiked into a composite buccal swab sample from healthy patients. This mock sample was extracted using the viral RNA isolation protocol described above, except that total nucleic acid was eluted in 100 μL. Negative control libraries were prepared from water. Illumina Nextera XT was used for library prepara-

tion: indexed libraries were generated using 16 cycles of PCR, and each sample was indexed with a unique barcode. Libraries were pooled equally based on molar concentration and sequenced on the Illumina HiSeq 2500 (100 or 150 bp paired-end reads) platform.

5.5.8 Hybrid Capture

Viral hybrid capture was performed as previously described [1] using two different probe sets. In one case, probes were created to target MuV and measles virus (MeV), and in one case, probes were created to target 356 species of viruses known to infect humans (V-All probe set [7]). Capture using V-All was used to enrich viral sequences primarily in samples in which we could not detect MuV, as well as in other samples. As described in Siddle et al. [7], the probe sets were designed to capture the diversity across all publicly available sequences on GenBank [11] for the relevant viruses.

5.5.9 Genome Assembly

We used viral-ngs v1.18.1 [28] to assemble reads form all sequencing runs. We used a set of MuV sequences (GenBank accessions: JX287389.1, FJ211586.1, AB000386.1, JF727652.1, AY685920.1, AB470486.1, GU980052.1, NC_002200.1, AF314558.1, AB823535.1, AF467767.2) to taxonomically filter these reads. We de novo assembled reads and scaffolded against the MuV genome with accession JX287389.1. We pooled data from all sequencing replicates of a sample, and then repeated this processes to obtain final genomes. Each time we ran viral-ngs, we set the 'assembly_min_length_fraction_of_reference' and 'assembly_min_unambig' parameters to 0.01.

We replaced deletions in the consensus genome coding regions with ambiguity ('N'). In one

sample, MuVs/Massachusetts.USA/11.16/5[G], with an insertion at position 3,903 (based on a full 15,384-nucleotide MuV genomes, e.g., accession JN012242.1) we removed a poorly-supported (<5 reads covering the site) extra 'A' in a homopolymer region.

5.5.10 Metagenomic Analysis

We used the V-All capture method on all samples from suspected mumps cases with a negative MuV PCR result (n=29). A subset of PCR-positive samples were also sequenced with this probe set (n=145) or without capture ('unbiased', n=111). We used the mock Enders strain MuV sample as a positive control on a sequencing run containing all PCR-negative samples, and used a water sample as a negative control. We obtained a partial mumps virus assembly using the viral-ngs method described above [28] in six PCR-negative samples, but observed high sequence similarity between some parts of these assemblies and the mock sample, which we considered evidence of contamination. Therefore, we prepared new sequencing libraries from all samples with evidence of other viruses and sequenced these replicates in the absence of the mock mumps sample. We required both replicates to contain reads matching any virus detected in the sample, and we found no evidence of other viruses in PCR-positive samples.

We used the metagenomic tool Kraken [29] to identify the source genera of the reads in each sample. We required total raw read count for any sample-genus pair to be twice (in practice, seven times) that in any negative control (water) from the same run. For any sample that had one or more pathogenic viral genera that passed this filter and had de-duplicated reads well distributed across the relevant viral genome, we attempted contig assembly: we filtered all sample

reads against all GenBank [11] nucleotide entries matching the identified species and then de novo assembled reads and scaffolded against the GenBank RefSeq [30] genome for the identified virus, using the viral-ngs [28] pipeline described above. We report all viruses identified via this method in Table E.2.

5.5.11 Criteria for Pooling Across Replicates

We prepared one or more sequencing libraries from each sample and attempted to sequence and assemble a genome from each of these replicates. We required a replicate to contain 3,000 unambiguous base calls for inclusion in the final genome assembly. This threshold was based on the maximum number of unambiguous bases (2,820) observed in negative controls (water samples) across all uncontaminated sequencing batches. One sequencing batch showed evidence of contamination: we were able to assemble 7,615 unambiguous MuV bases from a water sample, with a median coverage of $4\times$. For samples prepared in this batch only we implemented an additional requirement for including a replicate in pooling: the assembly must have a median coverage of at least $20\times$, five times the median coverage of the water sample.

5.5.12 Multiple Sequence Alignment of Genotype G Whole Genomes

We required a MuV genome to contain 11,538 unambiguous base calls (75% of the total genome with GenBank accession JN012242.1) for inclusion in the alignment of whole genome sequences. For two patients with samples taken at two time points (MuVs/Massachusetts.USA/19.16/5[G](1) and MuVs/Massachusetts.USA/19.16/5[G](2-20.16); MuVs/Massachusetts.USA/16.16/6[G](1) and

MuVs/Massachusetts.USA/16.16/6[G](2-17.16)), we only included the earlier sample in downstream analyses. The final alignment of whole genome sequences contains only samples belonging to genotype G; we did not include MuVs/Massachusetts.USA/24.17/5[K], which belongs to genotype K, in the alignment.

In this alignment, we also included 25 MuV genomes published on NCBI GenBank [11]. These comprise all of the sequences with organism 'Mumps rubulavirus' available as of September 2017 that meet the following criteria: sequence length of at least 14,000 nucleotides, belong to genotype G, sample collection year and country of origin reported in GenBank, no evidence of extensive virus passaging or modification (for vaccine development, for example). We aligned MuV genomes using MAFFT v7.221 [31] with default parameters.

5.5.13 Visualization of Coverage Depth Across Genomes

We plotted aggregate depth of coverage across the 200 samples whose genomes were included in the final alignment. For each sample, we aligned cleaned reads (as output by the depletion step of viral-ngs) to a reference genome (GenBank accession JX287389.1) with viral-ngs, using the novoalign aligner with options '-r Random -l 40 -g 40 -x 20 -t 100 -k'. Excluding duplicate reads, we calculated the depth of coverage at each nucleotide position in each sample. Then, we scaled the depth values within each sample so that each would have a mean depth of 1.0. We calculated the median depth taken across the samples for each nucleotide position, and the 20th and 80th percentiles, and plotted the mean of these metrics within a 200-nucleotide sliding window.

5.5.14 Analysis of Within- and Between-Sample Variants

We ran V-Phaser 2.0 via viral-ngs on all pooled reads mapping to a sample assembly to identify within-sample variants. To call a variant, we required a minimum of five forward and reverse reads, as well as no more than 10-fold strand bias, as previously described Gire et al. [12]. Sequences generated from the contaminated sequencing batch mentioned above (see Section 5.5.11). When analyzing variants of known contacts, we used pairs of samples designated as 'contact links', as described in Section 5.5.18 below. All within-host variant coordinates are as compared to the full 15,384 nucleotide genome (GenBank accession: JN012242.1).

Between-sample variants were called by comparing each final genome sequence to JX287385.1, the earlier of the two available whole genomes from the 2006 U.S. mumps outbreak. We adjusted coordinates to match those used in within-host variant analysis (GenBank accession: JN012242.1). We ignored all fully or partially ambiguous base calls, and excluded sequences that did not descend from the USA_2006 clade (Figure 5.1A) from this analysis. When examining amino acid changes in HN given vaccination status, we ignored sequences from patients with unknown vaccination history.

We used BEAST v1.8.4 [32] to perform both the site-specific and per-gene d_N/d_S analyses (Figure 5.3A–B). In both these analyses, we used the 225 MuV genotype G genomes included in the final alignment (200 sequences we generated plus 25 published sequences). The site-specific analysis implements the counting method described in Lemey et al. [33] on codon regions (CDS) of the MuV genome, excluding the portion of the V protein after the insertion site [34]. We partitioned

the CDS alignment into codon positions and used a separate HKY [35] substitution model and uncorrelated relaxed clock with log-normal distribution (UCLN) [36] on each partition. We used the trees generated when running BEAST on the full dataset (see Section 5.5.16) and ran 10 million MCMC steps to generate site-specific counts, sampling every 10,000 states.

For the per-gene analysis, we partitioned the alignment by gene (again excluding the ambiguous portion of the V protein) and ran BEAST as described in Park et al. [13]. We ran 200 million MCMC steps (sampling every 10,000 states) using a Bayesian Skyline tree prior [37] and an independent GY94 codon substitution model [38] for each gene partition. We used a HKY substitution model with Γ_4-distributed rate heterogeneity for the noncoding region [35, 39]) and parameterized all partitions with independent strict molecular clocks. For samples without exact dates, we used sampled tip dates [40].

5.5.15 Maximum Likelihood Estimation and Root-to-Tip Regression

We generated a maximum likelihood tree using the whole genome genotype G multiple sequence alignment. The tree was created using IQ-TREE v1.3.13 [41] with a GTR substitution model [42]. We rooted the tree on the oldest sequence in this dataset (GenBank accession: KF738113.1) in FigTree v1.4.2 [43].

To estimate root-to-tip distance of samples in the primary U.S. lineage, we subsetted the full genotype G alignment to include only samples descending from the USA_2006 clade (Figure 5.1A). We rooted this tree on the USA_2006 node and used TempEst v1.5 [44] to estimate distance from the root. We used scikit-learn in Python [45] to perform linear regression of distances on dates.

We also generated maximum likelihood trees using the SH gene only (full 316-nucleotide mRNA), HN (CDS only), F (CDS only), and a concatenation of the aforementioned SH, HN, and F regions. For each tree, we started with the whole genome genotype G alignment (225 sequences) and extracted the relevant region. We then removed any sequence with two or more consecutive ambiguous bases ('N's) in any of SH, HN, or F, leaving 209 sequences in each alignment. We used IQ-TREE v1.5.5 [41] with a GTR substitution model to generate maximum likelihood trees.

5.5.16 Molecular Dating Using BEAST

We used all 225 MuV genotype G genomes for molecular clock phylogenetics using BEAST v1.8.4 [32]. On the coding regions, we used the SRD06 substitution model, which partitions codons into two partitions (positions 1+2, position 3) and uses an HKY model [35] with gamma site heterogeneity. On the noncoding regions, we used a HKY substitution model also with gamma site heterogeneity. We again removed codon sites in the V protein after the insertion site [34] because of reading frame ambiguity in that region. For samples without exact dates, we used sampled tip dates [40].

We performed model selection as described in Metsky et al. [46]. We tested two clock models (a strict clock and an uncorrelated relaxed clock with log-normal distribution (UCLN) [36]) and three coalescent tree priors (a constant size population, an exponential growth population, and a Bayesian Skygrid tree prior with 20 groups [47]). We performed path-sampling (PS) and stepping-stone sampling (SS) on all six combinations of models and sampled for 100 path steps with a chain length of 2 million. The Skygrid tree prior in combination with a relaxed clock provided the best

model fit, as shown by the log Bayes factor when comparing this to other models (Table E.4), so we used these parameters for all other analyses. In all BEAST runs, we discarded the first 10% of states from each run as burn-in.

To obtain the tree and tMRCA estimates shown in Figure 5.1 and Table E.4, we ran BEAST with 200 million MCMC steps. We used Tracer v1.6.0 and TreeAnnotator v1.8.4 to extract tMRCA estimates and the maximum clade credibility (MCC) tree, respectively. We report the 95% highest posterior density intervals for selected nodes in Figure 5.1B and Table E.4.

5.5.17 Principal Components Analysis

The dataset for PCA consisted of all biallelic SNPs in the set of all 225 genotype G genomes. Missing data was imputed with the R package missMDA [48] and principal components were calculated with the R package FactoMineR [49]. Fourteen samples were discarded as outliers, based on visual inspection, leaving 212 samples in the final set.

5.5.18 Relationship Between Epidemiological and Genetic Data

We obtained detailed epidemiological data for samples shared by the MDPH from the Massachusetts Virtual Epidemiologic Network (MAVEN) surveillance system. From recorded fields and case notes, we defined two types of epidemiological links: 'contact links', individuals who were known contacts verified by local boards of health and/or other public health officials, and had symptom onset dates 7–33 days apart [50]; and 'shared activity links', individuals who participated in the same extra-curricular activity (i.e., a sports team or university club) or frequented a specific

residence or athletic facility. When analyzing the relationship between genetic and epidemiological data, we grouped both types of links.

We calculated pairwise genetic distance between all pairs of samples in the whole genome genotype G alignment. For each pair, the genetic distance score is s/n, where s is the number of unambiguous differing sites (both sequences must have an unambiguous base at the site, and the called bases must differ) and n is the number of sites at which both sequences have unambiguous base calls.

To calculate a multidemensional scaling (MDS) from the genomic distance matrix, we used the R package cmdscale [51]. To determine the ability of genetic distance to predict epidemiological linkage, we looked specifically within the II-outbreak (Figure 5.1A) clade, which is comprised of mostly Institution A, and the related community outbreak, cases. We constructed a receiver operating characteristic (ROC) curve using pairwise distance between II-outbreak cases as the predictor variable and presence or absence of an epidemiological link as the binary response variable.

5.5.19 Model of Mumps Transmission in a University Setting

We developed a stochastic model for mumps transmission accounting for the natural history of infection, vaccination status, and control measures implemented in response to the outbreak at Institution A. We used previous estimates of the effectiveness and waning rate of mumps vaccination [52], and of the vaccination status distribution of individuals on a university campus [53], to account for susceptibility to infection among the Institution A population ($N = 22,000$). Risk for mumps virus infection, given exposure, was scaled to time since receipt of the last vaccine dose,

yielding the hazard ratio $\xi_i = e^{\omega_0 \tau_i^{\omega_1}}$ for an individual i who received his/her last dose τ_i years previously, relative to an unvaccinated individual. For fitted values from [52], estimates were below one for individuals vaccinated since 1967, when the Jeryl Lynn vaccine was introduced. We plot the resulting susceptibility distribution in Figure E.4D.

Our stochastic model of mumps virus transmission included three stages after initial infection, the durations of which we inferred using data from previous clinical studies Figure E.4A–C. These included the gamma-distributed incubation period from infection to onset of mumps virus shedding in saliva [54]; the gamma-distributed period of latent infection from shedding onset to parotitis onset [54, 55]; and the log-normally distributed time from parotitis onset to the cessation of shedding [56]. For asymptomatic cases, we defined the total duration of shedding (γ) as the sum of independent random draws from the durations of shedding before and after parotitis onset, based on the lack of any reported difference in durations of shedding for symptomatic and asymptomatic cases [54]. To account for case isolation precautions implemented by Institution A, we modeled the removal of symptomatic individuals one day after onset of parotitis. In comparison to the 70% probability for symptoms given infection among unvaccinated individuals [57], we modeled the probability of symptoms given infection as uniformly distributed between 27.3% and 38.3% [52, 58].

Given the instantaneous hazard of infection for an as-yet uninfected individual i exposed to $I(t)$ infected individuals $\lambda_i(t) = \beta \xi_i I(t) N^{-1}$, the probability of evading infection over the course of a one-day simulated time step was $e^{-\lambda_i(t)}$. The per-contact transmission rate (β) was measured from the initial (pre-introduction) value of the effective reproductive number: $\beta = R_E(0)\bar{\gamma}\bar{\xi}^{-1}$.

5.5.20 Inferring Transmission Dynamics

The number of cases (71) and identification of multiple, distinct viral clades suggested limited permeation of mumps virus within Institution A after any introduction. We simulated dynamics of individual transmission chains to understand the epidemiological course of introduced viral lineages, and to infer values of $R_E(0)$ and the number of importations of mumps virus. We used the simulation model to sample from the distribution of the number of cases (X, including the index infection if symptomatic) resulting from a single introduction over a 1.5 year time course:

$$F\{x_i \mid R_E(0)\} = P[X = x_i \mid R_E(0)]$$

We resampled according to $f\{x_i \mid R_E(0)\}$ to define the distribution of the cumulative number of cases (Z) resulting from Y introductions, conditioned on $R_E(0)$:

$$g\{z_k \mid R_E(0), Y\} = P[Z = z_k = \sum_{i=1}^{y_i} x_i \mid R_E(0), Y = y_j]$$

For simulations where $Z \geq 66$, we drew $k = 66$ cases at random to determine the number of distinct lineages (S, defined by the index infection) expected to be present within such a sample. The probability of obtaining 66 sequences, and observing $S = s_m$ lineages among them, is

$$h\{s_m \mid R_E(0), Y, K = 66\} = P[S = s_m \mid R_E(0), Y = y_j, K = 66] \times P[Z \geq 66 \mid R_E(0), Y]$$

The posterior density of our model also accounted for the probability of observing 71 symptomatic cases in total. Defined in terms of the number of introductions and the initial reproductive number, the model posterior was proportional to

$$h\{4 \mid R_E(0), Y, K = 66\} \times g\{71 \mid R_E(0), Y\}$$

We measured this probability from 100,000 iterates for each pairing of $R_E(0) \in 0.10, 0.11, \ldots, 2.50$ and $Y \in 1, 2, \ldots, 200$.

5.5.21 Transmission Reconstruction Using Outbreaker

We used the R package outbreaker [59] to reconstruct transmission for samples included in clade II-outbreak. We infer the distribution of the generation interval length using data from ten cases in our dataset with known exposure sources. A gamma distribution fitted by maximum likelihood (Figure E.4E) recovers mean and dispersion estimates nearly identical to those reported in earlier mumps outbreaks [60]. We ran outbreaker six times in parallel, each with 1 million MCMC steps, and discarded the first 10% of states as the burn-in.

5.5.22 SH and HN Multiple Sequence Alignment

To analyze all published SH and HN MuV sequences, we searched NBCI GenBank in July 2017 for all nucleotide sequences with organism 'Mumps rubulavirus'. We performed a pairwise alignment between each sequence result and a reference genome (GenBank accession: JX287389.1) using

MAFFT [31] with parameters: '--localpair --maxiterate 1000 --preservecase'. We then extracted the SH sequence from each pairwise alignment, removing all sequences without the full 316-nucleotide region and all sequences with an insertion or deletion ('indel') relative to the reference. We then used MAFFT with parameters '--localpair --maxiterate 1000 --retree 2 --preservecase' to create a multiple sequence alignment of the extracted SH gene sequences and removed any sequences with indels in this final alignment. We repeated the same process for the HN region, requiring the full 1749-nucleotide coding region.

In both the SH and HN alignments, we removed sequences from vaccine strains (i.e., genotype N, or another genotype marked as "(VAC)" or "vaccine"). We also removed sequences with GenBank records indicating extensive passaging. In the SH alignment only, we removed sequences with no reported collection date or country of origin, as these data are required for phylogeographic analyses. In samples with a collection decade (e.g., 1970s) but not a specific year, we assigned the first year of the decade; in samples with only a collection year, we assigned a decimal year of $year + 0.5$ (e.g., 1970.5); in samples with year and month but no day, we used the day halfway through the given month (e.g., March 2015 becomes 15 March 2015) to calculate the decimal year; and in samples with an epidemiological week but no specific day, we approximated the decimal year as $year + (epi\ week/52)$, except samples collected in epidemiological week 52 were relabeled as week 51.999 to avoid confusion. In the HN alignment, we simply used the reported year as the date (i.e., 15 March 2015 becomes 2015), and did not remove samples without date information.

In both the SH and HN alignments, we relabeled outdated genotypes (M, E, and any subgenotypes) and constructed a maximum likelihood tree (using IQ-TREE with a GTR substitution

model, as described in Section 5.5.15) to assign a genotype if one was not reported on GenBank. We preserved genotypes designated as 'Unclassified' [61].

To each alignment, we appended all SH or HN sequences from individual patients generated in this study, except those with two or more consecutive ambiguous bases ('N's) in the SH or HN region.

5.5.23 SH Phylogeographic Analysis

To perform phylogenetic and phylogeographic analyses of the SH gene sequence, we first sampled trees using BEAST v1.8.4 [32]. We used constant size population and strict clock models. Other demographic and clock models would have likely provided a better fit to the data, but it appeared that using them (especially the use of a relaxed clock) would have made it impractical to achieve convergence and sufficient sampling of trees and all parameters on the full dataset. We used the HKY substitution model [35] with four rate categories and no codon partitioning. We ran BEAST in four replicates, each for 500 million states with sampling every 50,000 states, and removed the first 150 million states as burn-in. We verified convergence of all parameters across the four replicates. Finally, we combined the four replicates using LogCombiner and resampled to obtain the final sampled states. On the resulting MCC tree (Figure 5.4A), we colored tips by 15 world regions shown in Figure E.6A. Using these sampled states, we computed a kernel density estimate of the probability distribution of the tMRCA for the 11 genotypes in this dataset (Figure 5.4B) with ggplot2 [62].

This dataset has large temporal and spatial sampling biases that affect estimates of migration and, to a lesser extent, ancestry. Using a structured coalescent to model migration and co-

alescent processes could alleviate these biases, but might face practical limitations on this large dataset [63]. Here, we used resampling on the input sequences to construct distributions of estimates. To perform this resampling, we focused on only samples that were collected both within a window of time and from a geographic region with sufficient sampling. Namely, we considered only sequences sampled in 2010 or afterward and collapsed the locations used earlier to just four regions: United States (consisting of only samples from the United States), Europe (consisting of samples whose location was labeled as Eastern Europe, Northern Europe, Southern Europe, Western Europe, or the United Kingdom), East Asia (consisting of samples whose location was labeled as Eastern Asia or Japan), and South/Southeast Asia (consisting of samples whose location was labeled as Southeast Asia or Southern Asia). We ignored samples from the five locations not used here (Canada; Caribbean, Central America, and South America; Middle and Eastern Africa; Northern Africa; Middle East). Then, we randomly sampled 10 sequences (without replacement) from each region for each year (i.e., 2010–2011, 2011–2012, and so on). We resampled the input sequences with this strategy 100 times.

Note that sampling biases affect this resampling strategy as well. Notably, several years in the East Asia and the South/Southeast Asia regions have fewer than 10 sequences (sometimes, zero) available for resampling, which may lead to an underestimate on the relative rates of migration involving these regions. Moreover, the high sequence similarity between SH gene sequences from United States and Europe (Figure 5.4D) may bias upward the distributions on the relative rates of migration between these regions and on the proportion of ancestry shared between them (Figure E.6) (e.g., if there have been few true migrations between these regions, but the gene has not accu-

mulated substitutions in the time between those migrations).

For each of the 100 resamplings of the input sequences, we ran BEAST to sample trees, as described above, for 100 million states sampling every 10,000 states; we removed 10 million states as burn-in and resampled to obtain 1,000 sampled trees. Then, for each of the 100 samplings of trees, we then performed a phylogeographic analysis with $4^2 - 4 = 12$ rates for 10 million states sampling every 1,000 states; we removed 1 million states as burn-in and resampled both the complete Markov jump history and the trees with ancestral locations every 10,000 states. For each of these 100 samplings of trees we then calculated the MCC tree using TreeAnnotator, and then calculated the final MCC tree from among the 100 options.

When plotting probability distributions showing the fraction of migrations between to each region from each other (Figure 5.4C), we calculated the mean of this fraction across all the sampled states for each run to produce a point estimate for each resampling of the input sequences, and show the distribution of these means across the 100 runs. Similarly, we used Posterior Analysis of Coalescent Trees (PACT) [64] to estimate tip ancestry on the sampled trees from each of the 100 runs; the proportion of ancestry we plot between a pair of locations in a time window is the mean across the 100 runs of the mean proportion for that pair in that time window from each run. We calculated the Bayes factors on migration routes between the four regions in Figure E.6C by combining the sampled indicator variables across all 100 runs to compute the posterior odds, and took the prior odds to be 1/3 (since N=4). We calculated the pointwise percentile bands in Figure E.6D from the mean proportions in each run across the 100 runs (i.e., they are percentiles across the resamplings of the input sequences).

5.6 References

[1] Matranga, C. B., Andersen, K. G., Winnicki, S., et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol*, 15(519):519, 2014.

[2] Centers for Disease Control and Prevention. Mumps Cases and Outbreaks, 2018. URL https://www.cdc.gov/mumps/outbreaks.html.

[3] Centers for Disease Control and Prevention. FastStats - Immunization, 2017. URL https://www.cdc.gov/nchs/fastats/immunize.htm.

[4] Dayan, G. H., Quinlisk, M. P., Parker, A. A., et al. Recent resurgence of mumps in the United States. *New England Journal of Medicine*, 358(15):1580–1589, 2008.

[5] Centers for Disease Control and Prevention. National Notifiable Diseases Surveillance System, 2017. URL https://data.cdc.gov/browse?category=NNDSS.

[6] Barskey, A. E., Schulte, C., Rosen, J. B., et al. Mumps outbreak in Orthodox Jewish communities in the United States. *New England Journal of Medicine*, 367(18):1704–1713, 2012.

[7] Siddle, K. J., Metsky, H. C., Gladden-Young, A., et al. Capturing diverse microbial sequence with comprehensive and scalable probe design. *bioRxiv*, 2018. doi: 10.1101/279570. URL https://www.biorxiv.org/content/early/2018/03/12/279570.

[8] Barrabeig, I., Costa, J., Rovira, A., et al. Viral etiology of mumps-like illnesses in suspected mumps cases reported in Catalonia, Spain. *Human vaccines & immunotherapeutics*, 11(1):282–287, 2015.

[9] Davidkin, I., Jokinen, S., Paananen, A., Leinikki, P., and Peltola, H. Etiology of mumps-like illnesses in children and adolescents vaccinated for measles, mumps, and rubella. *The Journal of Infectious Diseases*, 191(5):719–723, 2005.

[10] Centers for Disease Control and Prevention. Influenza & Parotitis: Question & Answers for Health Care Providers, 2016. URL https://www.cdc.gov/flu/about/season/questions-answers-parotitis.htm.

[11] Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. GenBank. *Nucleic acids research*, 44(Database issue):D67–72, 2016.

[12] Gire, S. K., Goba, A., Andersen, K. G., et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[13] Park, D. J., Dudas, G., Wohl, S., et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526, 2015.

[14] Didelot, X., Gardy, J., and Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol*, 31(7):1869–1879, 2014.

[15] Wolinsky, J. S., Waxham, M. N., and Server, A. C. Protective effects of glycoprotein-specific monoclonal antibodies on the course of experimental mumps virus meningoencephalitis. *Journal of Virology*, 53(3):727–734, 1985.

[16] Kövamees, J., Rydbeck, R., Orvell, C., and Norrby, E. Hemagglutinin-neuraminidase (HN) amino acid alterations in neutralization escape mutants of Kilham mumps virus. *Virus research*, 17(2):119–129, 1990.

[17] Orvell, C., Alsheikhly, A. R., Kalantari, M., and Johansson, B. Characterization of genotype-specific epitopes of the HN protein of mumps virus. *J Gen Virol*, 78 (Pt 12):3187–3193, 1997.

[18] Cusi, M. G., Fischer, S., Sedlmeier, R., et al. Localization of a new neutralizing epitope on the mumps virus hemagglutinin-neuraminidase protein. *Virus research*, 74(1-2):133–137, 2001.

[19] Kulkarni-Kale, U., Ojha, J., Manjari, G. S., et al. Mapping antigenic diversity and strain specificity of mumps virus: A bioinformatics approach. *Virology*, 359(2):436–446, 2007.

[20] Tanabayashi, K., Takeuchi, K., Hishiyama, M., et al. Nucleotide sequence of the leader and nucleocapsid protein gene of mumps virus and epitope mapping with the in vitro expressed nucleocapsid protein. *Virology*, 177(1):124–130, 1990.

[21] Rubin, S. A., Qi, L., Audet, S. A., et al. Antibody induced by immunization with the Jeryl Lynn mumps vaccine strain effectively neutralizes a heterologous wild-type mumps virus associated with a large outbreak. *The Journal of Infectious Diseases*, 198(4):508–515, 2008.

[22] Yeo, R. P., Afzal, M. A., Forsey, T., and Rima, B. K. Identification of a new mumps virus lineage by nucleotide sequence analysis of the SH gene of ten different strains. *Archives of virology*, 128(3-4):371–377, 1993.

[23] Jin, L., Rima, B., Brown, D., et al. Proposal for genetic characterisation of wild-type mumps strains: preliminary standardisation of the nomenclature. *Archives of virology*, 150(9):1903–1909, 2005.

[24] Jin, L., Örvell, C., Myers, R., et al. Genomic diversity of mumps virus and global distribution of the 12 genotypes. *Reviews in medical virology*, 25(2):85–101, 2015.

[25] Marin, M., Marlow, M., Moore, K. L., and Patel, M. Recommendation of the Advisory Committee on Immunization Practices for Use of a Third Dose of Mumps Virus–Containing Vaccine in Persons at Increased Risk for Mumps During an Outbreak. *MMWR. Morbidity and mortality weekly report*, 67(1):33–38, 2018.

[26] Centers for Disease Control and Prevention. Real-time (TaqMan) RT-PCR Assay for the Detection of Mumps Virus RNA in Clinical Samples , 2010. URL https://www.cdc.gov/mumps/downloads/lab-rt-pcr-assay-detect.pdf.

[27] Van Camp, G., Chapelle, S., and De Wachter, R. Amplification and sequencing of variable regions in bacterial 23S ribosomal RNA genes with conserved primer sequences. *Current microbiology*, 27(3):147–151, 1993.

[28] Tomkins-Tinch, C., Ye, S., Metsky, H., et al. viral-ngs, 2016. URL doi:10.5281/zenodo.200428.

[29] Wood, D. E. and Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

[30] O'Leary, N. A., Wright, M. W., Brister, J. R., et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–45, 2016.

[31] Katoh, K. and Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, 2013.

[32] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*, 29(8):1969–1973, 2012.

[33] Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L., and Suchard, M. A. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, 28(24):3248–3256, 2012.

[34] Paterson, R. G. and Lamb, R. A. RNA Editing by G-Nucleotide Insertion in Mumps Virus P-Gene mRNA Transcripts. *Journal of Virology*, 64(9):4137–4145, 1990.

[35] Hasegawa, M., Kishino, H., and Yano, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[36] Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, 4(5):e88, 2006.

[37] Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*, 22(5):1185–1192, 2005.

[38] Goldman, N. and Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.

[39] Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, 1994.

[40] Shapiro, B., Ho, S. Y. W., Drummond, A. J., et al. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol Biol Evol*, 28(2):879–887, 2011.

[41] Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015.

[42] Tavare, S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17, 1986.

[43] Rambaut, A. FigTree, 2014. URL http://tree.bio.ed.ac.uk/software/figtree/.

[44] Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*, 2(1):vew007, 2016.

[45] Abraham, A., Pedregosa, F., Eickenberg, M., et al. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

[46] Metsky, H. C., Matranga, C. B., Wohl, S., et al. Zika virus evolution and spread in the Americas. *Nature*, 546(7658):411–415, 2017.

[47] Gill, M. S., Lemey, P., Faria, N. R., et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724, 2013.

[48] Josse, J. and Husson, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *J. Stat. Softw.*, 70(1):1–31, 2016.

[49] Lê, S., Josse, J., and Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.*, 2008.

[50] Centers for Disease Control and Prevention. Mumps: For Healthcare Providers, 2018. URL https://www.cdc.gov/mumps/hcp.html.

[51] R Core Team. R: A Language and Environment for Statistical Computing, 2016. URL https://www.R-project.org/.

[52] Lewnard, J. A. and Grad, Y. H. Vaccine waning and mumps re-emergence in the United States. *Science Translational Medicine*, 2018.

[53] Cardemil, C. V., Dahl, R. M., James, L., et al. Effectiveness of a Third Dose of MMR Vaccine for Mumps Outbreak Control. *New England Journal of Medicine*, 377(10):947–956, 2017.

[54] HENLE, G. and HENLE, W. Isolation of mumps virus from human beings with induced apparent or inapparent infections. *The Journal of experimental medicine*, 88(2):223–232, 1948.

[55] Ennis, F. A. and Jackson, D. Isolation of virus during the incubation period of mumps infection. *The Journal of pediatrics*, 72(4):536–537, 1968.

[56] Polgreen, P. M., Bohnett, L. C., Cavanaugh, J. E., et al. The duration of mumps virus shedding after the onset of symptoms. *Clinical Infectious Diseases*, 46(9):1447–1449, 2008.

[57] Galazka, A. M., Robertson, S. E., and Kraigher, A. Mumps and mumps vaccine: a global review. *Bulletin of the World Health Organization*, 77(1):3–14, 1999.

[58] Fanoy, E. B., Cremer, J., Ferreira, J. A., et al. Transmission of mumps virus from mumps-vaccinated individuals to close contacts. *Vaccine*, 29(51):9551–9556, 2011.

[59] Jombart, T., Cori, A., Didelot, X., et al. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLOS Computational Biology*, 10(1):e1003457, 2014.

[60] Vink, M. A., Bootsma, M. C. J., and Wallinga, Jacco. Serial intervals of respiratory infectious diseases: a systematic review and analysis. *American journal of epidemiology*, 180(9):865–875, 2014.

[61] WHO. Mumps virus nomenclature update: 2012, 2012.

[62] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[63] Müller, N. F., Rasmussen, D. A., and Stadler, T. The Structured Coalescent and Its Approximations. *Molecular Biology and Evolution*, 34(11):2970–2981, 2017.

[64] Bedford, T. PACT. URL https://github.com/trvrb/PACT.

Chapter 6

Conclusion

The work presented in this dissertation represents several examples of using genomics to respond to and understand viral outbreaks, through the studies of Ebola, Zika, and mumps viruses. In each case, we used genomic variation to improve our understanding of the origin of the virus, its spread during the outbreak, and/or the possible functional implications of specific variant sites. These fundamental aims, as laid out in the introduction, were common across chapters, but each study presented unique challenges and goals.

In Chapter 2, we found there was limited movement of Ebola virus (EBOV) across the Sierra Leonean border, made the case for purifying selection in EBOV during the course of the epidemic, and identified two pieces of evidence for host effects on the viral genome. These findings not only answered questions about the outbreak, but also suggested new hypotheses that may aid in combating future EBOV outbreaks. For example, we identified many new variants shared across EBOV genomes. As shown by a recent study [1] on the mutation defining the SL3 EBOV lineage [2], mutations identified in this type of large sequencing study, especially ones deemed of particular in-

terest by phylogenetic relationships, can focus experimental investigation and lead to important advances in our understanding of the virus.

Our work on EBOV in Sierra Leone also shows that within-host variants are likely passed between individuals infected with EBOV, and may be useful in understanding transmission. We used this observation in Chapter 3, in which I explain how we used both between- and within-host variants to reconstruct the EBOV transmission tree during the outbreak in Nigeria. In addition to highlighting the ability of genomic data to reconstruct transmission patterns, variation in the virus allowed us to place the outbreak in Nigeria in the context of the larger EBOV epidemic. We showed that the virus likely entered the country from Liberia, and that EBOV did not spread from Nigeria to other countries, thus confirming the success of the public health response. The data presented in this chapter also underscore the need for better understanding of within-host dynamics of EBOV, which could aid in both transmission reconstruction and identifying variants of potential functional or therapeutic interest. This is already an area of active interest [3], but further experimental studies are needed. Although the high risk of working with EBOV, a BSL-4 virus, makes this a challenging endeavor, it is an important step in gaining a more complete picture of this deadly virus.

When studying Zika virus (ZIKV), we were most interested in understanding viral entry into the Americas, and found evidence for ZIKV in four different countries several months before it was diagnosed. Genomic data made it clear that many ZIKV cases were likely missed by traditional diagnostics, and that this is a key area of improvement for combating the spread of the virus. Future studies should not only focus on improving our ability to detect ZIKV (an area of research already

rapidly evolving [4–6]), but also on methods that can be used to diagnose viral infection without a priori knowledge of the infecting pathogen. In Chapter 4, I also describe our investigation of within-host variation in ZIKV using two different sequencing methods. This analysis demonstrates the need for improved sequencing methods for low-titer viruses, as well as better ways to distinguish true within-host variants from ones caused by sequencing error or contamination.

While genomic analysis of EBOV and ZIKV produced important conclusions related to country-level viral origins and spread, the analysis of mumps virus (MuV) described in Chapter 5 highlights the ability of genomic data to inform public health on a variety of geographic scales. In this chapter, genomic data suggest ongoing MuV transmission in the United States, a conclusion not obvious from the pattern of sporadic outbreaks throughout the country. Additionally, we were able to link the outbreaks in two very different communities within Massachusetts, suggesting transmission of the virus outside of universities (where most cases were identified) into the local community. Despite an apparent lack of informative within-host variation in MuV, we were able to reconstruct transmission at high resolution. We also addressed the hypothesis of vaccine escape during the outbreak, and found no strong evidence of this occurring. That said, we identified two amino acid sites that differ between the dataset we generated and published genomes from a large outbreak in the United States in 2006. These sites should be carefully investigated to determine if they affect virus neutralization by vaccine-induced antibodies.

This dissertation covers many topics in viral genomics and the potential of sequencing to inform public health, yet one theme that emerges in all chapters is the potential of using variation — including within-host variation — to inform transmission analysis. Throughout, I point out the

need for improved within-host variant identification, and the necessity of rigorous methods that can harness this data to improve transmission reconstruction. A number of recent studies, including our ZIKV study, have addressed validation of iSNVs across replicate studies [7, 8] and methods for improving variant identification [9]. Sequencing samples with known variant frequencies (created by mixing different virus strains or synthetic oligonucleotides at various frequencies) could help generalize these methods. Sequencing at various read depths would also be important, since I have observed a clear correlation between depth and iSNV detection. Finally, technical replicates at each stage of sequencing preparation could help determine the primary source of errors (i.e., random sequencing error [10] or contamination during a particular process), potentially providing a tool for improving sequencing methods [9].

Of course, some viruses do not have meaningful within-host variation, as we saw in Chapter 5. In these cases, between-host variation may be sufficient for understanding transmission, or additional data may be required. For example, it may be possible to take advantage of metagenomic sequencing to bolster transmission analysis; other organisms may be transmitted with a viral pathogen and have the potential to provide valuable information about the individuals involved. This type of approach will rely heavily on the previous and ongoing work in the microbiome field, and will require careful consideration of metagenomic techniques to filter out background and contaminants from the sequencing data.

If we can confidently detect iSNVs, it is important that we know how to use them. This entails improving our understanding of virus-specific within-host dynamics and developing new computational methods that use within-host variation in transmission reconstruction. These aims are

connected: an accurate model of within-host viral dynamics is essential to building computational methods that make estimates from these models. As discussed briefly in the introductory chapter, within-host dynamics have been explored in more detail in some pathogens already (e.g., HIV [11]), and within-host dynamics have been incorporated into recent recent transmission reconstruction methods [12–14]. Most, however, assume a bottleneck size (the number of viruses transmitted from one individual to another) of one, and relaxing this assumption may be essential to proper transmission reconstruction of viruses like Ebola, which have been hypothesized to have a substantially large bottleneck [15, 16].

Finally, we also need to consider how these methods could and should be used in real time. During an ongoing outbreak, there may be missing cases or only partial transmission chains, and it is important to consider how transmission reconstruction methods account for missing data, and also how they should be optimized to answer key questions during an outbreak. These important questions include identification of super-spreaders — individuals who transmitted a virus to an exceptional number of secondary cases — and identification of groups or demographics most likely to be involved in pathogen transmission. For example, during the MuV outbreak, sports teams were widely considered to be a hotbed of MuV transmission and infection. Determining the individuals likely to be most affected by an outbreak can help focus monitoring and containment methods.

It is important to note that genomic data, even without iSNVs, has already been shown to greatly improve our understanding of a number of viral outbreaks. While continued efforts to understand within-host variation and dynamics may increase our ability to draw conclusions from se-

quencing data, widespread adoption of sequencing for pathogen surveillance is of equal, or greater, importance. Throughout my work on EBOV, ZIKV, and MuV, I have seen how conclusions drawn from genomic data can have a direct impact on our understanding of a viral outbreak and can be used to inform public health. Additionally, these projects suggest the importance of combining genomic and epidemiological data to most effectively trace and address outbreaks. In Chapter 5, for example, information about institutions most affected by MuV allowed us to explore spread into and within these specific communities, and epidemiological information about individuals affected will be essential to achieving the transmission-related goals (such as identifying demographics prone to viral transmission) stated above.

Genomic epidemiology and surveillance, despite mounting evidence of its value, remains challenging. Issues related to sample transfer, research approvals, and metadata were common across the EBOV, ZIKV, and MuV projects, highlighting the need for a better framework for real-time genomic surveillance everywhere. Sequencing technology has improved enough to make real-time genomic sequencing possible, and building the systems to facilitate sample collection, transfer, and sequencing are the necessary next step. Improving the experimental, computational, and logistical methods for genomic analysis of viral outbreaks will continue to expand our ability to inform public health measures and respond to viral disease outbreaks.

References

[1] Diehl, W. E., Lin, A. E., Grubaugh, N. D., et al. Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic. *Cell*, 167(4):1088–1098.e6, 2016.

[2] Gire, S. K., Goba, A., Andersen, K. G., et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[3] Ni, M., Chen, C., Qian, J., et al. Intra-host dynamics of Ebola virus during 2014. *Nature microbiology*, 1:16151, 2016.

[4] Gootenberg, J. S., Abudayyeh, O. O., Lee, J. W., et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science*, 356(6336):438–442, 2017.

[5] Bosch, I., de Puig, H., Hiley, M., et al. Rapid antigen tests for dengue virus serotypes and Zika virus in patient serum. *Science Translational Medicine*, 9(409):eaan1589, 2017.

[6] Pardee, K., Green, A. A., Takahashi, M. K., et al. Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell*, 165(5):1255–1266, 2016.

[7] Watson, S. J., Welkers, M. R. A., Depledge, D. P., et al. Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120205, 2013.

[8] Poon, L. L. M., Song, T., Rosenfeld, R., et al. Quantifying influenza virus diversity and transmission in humans. *Bioinformatics*, 25(16):2078–2079, 2009.

[9] Posada-Cespedes, S., Seifert, D., and Beerenwinkel, N. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus research*, 239:17–32, 2016.

[10] Nakamura, K., Oshima, T., Morimoto, T., et al. Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, 39(13):e90–e90, 2011.

[11] Lemey, P., Rambaut, A., and Pybus, O. G. HIV evolutionary dynamics within and among hosts. *AIDS reviews*, 8(3):125–140, 2006.

[12] Hall, M., Woolhouse, M., and Rambaut, A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology*, 11(12): e1004613, 2015.

[13] Worby, C. J., O'Neill, P. D., Kypraios, T., et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics*, 10(1): 395–417, 2016.

[14] Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., and Wallinga, Jacco. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5):e1005495, 2017.

[15] Park, D. J., Dudas, G., Wohl, S., et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526, 2015.

[16] Emmett, K. J., Lee, A., Khiabanian, H., and Rabadan, R. High-resolution Genomic Surveillance of 2014 Ebolavirus Using Shared Subclonal Variants. *PLoS Currents*, 7:–, 2015.

Appendix A

Supplemental Material for Chapter 1

A.1 Supplemental Figures and Tables

Table A.1: Software and online tools for genomic analysis, including assembly and alignment depicted in Figure 1.1. See Chapter 1 for indicated references.

Table A.1: Continued

Category	Primary Function(s)	Software name	Input requirements	Description	Reference
Assembly and alignment	Alignment manipulation	Samtools	Sequence alignment	• Sorts, merges, and indexes alignments • Retrieves reads in any region quickly (variant calling)	[48]
	Assembly	ABySS	High-throughput sequencing reads	• De novo assembly of short reads • Parallel processing available for assembly of large genomes	[149]
	Assembly	SOAPdenovo	High-throughput sequencing reads	• De novo assembly of short reads, optimized for Illumina reads	[150]
	Assembly	SPAdes	High-throughput sequencing reads	• De novo assembly of short reads • Designed for single cell and multicell bacterial datasets; not suitable for large genomes • Best for short, high quality reads where indels are rare	[151]
	Assembly	Trinity	High-throughput sequencing reads	• De novo assembly of Illumina RNA-seq reads	[152]
	Assembly	Velvet	High-throughput sequencing reads	• De novo assembly of short reads • Suitable for both prokaryotic and mammalian genomes	[153]
	Global alignment	MAFFT	Sequences to be aligned	• High accuracy multiple sequence alignment • Suitable for long sequences or sequences with large gaps or variable loop regions	[154,155]
	Global alignment	MUSCLE	Sequences to be aligned	• Rapid multiple sequence alignment	[156]
	Local alignment	BLAST	Reference sequence, query sequence	• Finds similar regions between sequences • Can be used to remove human reads from metagenomics datasets • Online interface available	[157]
	Local alignment	Lastal	Reference sequence, query sequence	• Finds similar regions between sequences • Deals well with repeat-rich sequences (e.g. whole genomes) and mismatches/gaps • Online interface available	[164]
	Quality scoring	FastQC	High-throughput sequencing reads	• Quality control tool for high-throughput sequencing reads	[165]
	Read manipulation	Picard	High-throughput sequencing reads	• Set of command-line tools for manipulating high-throughput sequencing reads	[172]
	Recombination screening	RDP4	Sequence alignment	• Detects and analyzes recombination and reassortment in aligned DNA sequences • Uses many different recombination detection methods	[46]
	Short read alignment	Bowtie	High-throughput sequencing reads, reference genome	• Fast, memory-efficient mapping of short reads onto a reference genome	[158]
	Short read alignment	BWA	High-throughput sequencing reads, reference genome	• Mapping of low-divergence sequences onto a reference genome	[159]
	Short read alignment	Mosaik	High-throughput sequencing reads, reference genome	• Accurate mapping of short reads onto a reference genome • Suitable for sequences with mismatches or short indels	[160]
	Short read alignment	NovoAlign	High-throughput sequencing reads, reference genome	• High-sensitvity mapping of short reads onto a reference genome • Slower but more sensitive than most other aligners	[171]
	Short read alignment	SNAP	High-throughput sequencing reads, reference genome	• Fast mapping of short reads onto a reference genome • Faster than BWA, Novoalign	[161]
Variant calling and annotation	Variant annotation	Annovar	List of variants, reference nucleotides	• Genome, region, and database-based variant annotation	[51]
	Variant annotation	Ensembl Variant Effect Predictor	List of variants, reference nucleotides	• Genome, region, and database-based variant annotation • Online interface available	[50]
	Variant annotation	SnpEff	List of variants, reference nucleotides	• Genome, region, and database-based variant annotation	[49]
	Variant calling	FreeBayes	High-throughput sequencing reads aligned to reference	• Haplotype-based variant calling for high-throughput sequencing data • Detects SNPs, indels, multi-base mismatches, complex variants	[162]
	Variant calling	GATK	Sequence alignment	• A set of tools for analyzing high-throughput sequencing reads • Especially good for variant calling (SNPs, short indels) and genotyping	[47]
	Variant calling	Platypus	High-throughput sequencing reads aligned to reference	• Haplotype-based variant calling for high-throughput sequencing data • Detects SNPs, indels, multi-base mismatches, replacements, large deletions	[163]
	Variant calling	VarScan	High-throughput sequencing reads aligned to reference	• Variant calling for high-throughput sequencing data, including within-host variant calling	[53]
	Variant calling	V-Phaser 2	High-throughput sequencing reads aligned to reference	• Variant calling for viral popualations	[52]

Table A.1: Continued

Category	Primary Function(s)	Software name	Input requirements	Description	Reference
Metagenomic analysis	Human read depletion	BMTagger	High-throughput sequencing reads, human reference genome	• Removes human reads from metagenomics datasets	[166]
	Interactive taxonomic analysis	MEGAN	High-throughput sequencing reads aligned to reference	• Sequencing reads must be aligned by BLAST or DIAMOND prior to running MEGAN • Processes BLAST/DIAMOND result into a file optimized for statistical and graphical analysis	[40]
	Taxonomic analysis	Kraken	High-throughput sequencing reads or sequences	• Fast assignment of taxonomic labels to short DNA sequences • Faster than MetaPhlAn but can be lower specificity	[41]
	Taxonomic analysis	MetaPhlAn	High-throughput sequencing reads	• Assignment of taxonomic labels to short DNA sequences • Fast, highly specific, but potentially low sensitivity in some environments	[42]
Phylogenetic analysis	Phylogenetic analysis	BEAST	Sequence alignment, sampling dates, nucleotide substitution model	• Bayesian analysis of molecular sequences using MCMC • Used for reconstructing phylogenies and evolutionary hypothesis testing	[70]
	Substitution rate estimation	TipDate	Sequence alignment, sampling dates	• Estimates the substituiton rate using maximum likelihood • Estimates date of most recent common ancestor of the sequences • Online interface available	[68]
	Tree visualization	FigTree	Phylogenetic tree file	• Application for viewing phylogenies and producing high-quality figures • Optimized for reading BEAST output and annotations	[168]
	Tree-building	MrBayes	Sequence alignment, nucleotide subsitution model	• Bayesian analysis of molecular sequences using MCMC	[58]
	Tree-building	PhyML	Sequence alignment, nucleotide subsitution model	• Phylogeny reconstruction using maximum likelihood • Simple, accurate, fast • Online interface available	[56]
	Tree-building	RAxML	Sequence alignment, nucleotide subsitution model	• Phylogeny reconstruction using maximum likelihood • Developed for handling large datasets with relatively low memory consumption	[57]
Transmission reconstruction	Transmission Reconstruction	outbreaker	Sequence alignment, sampling dates, generation time distribution	• Reconstructs disease outbreak transmission using a Bayesian approach • Can be used to simulate an outbreak and pathogen evolution	[96,100]
Evolutionary analysis	Selection analysis	HyPhy	Sequence alignment, nucleotide subsitution model	• Package for maximum likelihood-based evolutionary analysis • Tools include: identification of sites under selection, recombination screening • Online interface available (Datamonkey)	[123,126]
	Selection analysis	PAML	Sequence alignment, nucleotide subsitution model	• Package for maximum likelihood analysis of protein and DNA sequences • Tools include: identification of sites under selection, substitution rate calculation	[125]
	Synonymous constraint analysis	FRESCo	Sequence alignment, phylogenetic tree file	• Identifies regions with excess synonymous constraint in short, deep alignments	[124]
Pipelines and multi-tool software	Alignment, visualization, tree-building	CLC Genomics Workbench	Sequencing reads or alignment	• Software for analyzing high-throughput sequencing data (CLC Main Workbench for Sanger) • Includes assembly, read mapping, alignment, variant calling, tree-building • Includes tools for transcriptomic, epigenomic, and RNA structure analysis	[167]
	Alignment, visualization, tree-building	Geneious	Sequencing reads or alignment	• Software for analyzing both high-throughput and Sanger sequencing data • Includes assembly, read mapping, alignment, variant calling, tree-building	[169]
	Assembly and alignment pipeline	viral-ngs	High-throughput sequencing reads	• Pipeline for assembly and alignment of viral sequence reads • Includes read processing, de novo assmebly, intrahost variant analysis • Online interface available (DNAnexus)	[173]
	Metagenomics pipleline	PathSeq	High-throughput sequencing reads	• Pipeline for metagenomic analysis of human tissue samples • Best when the majority of reads are host reads	[44]
	Metagenomics pipleline	SURPI	High-throughput sequencing reads	• Pipeline for fast metagenomic analysis of clinical samples	[43]
	Outbreak visualization	Nextflu	Viral genome sequences	• Online interface and pipeline for real-time tracking of virus evolution in humans • Visualization is available online for seasonal influenza and Ebola virus • Source code for running the pipeline is available online	[170]

Appendix B

Supplemental Material for Chapter 2

B.1 Supplemental Figures and Tables

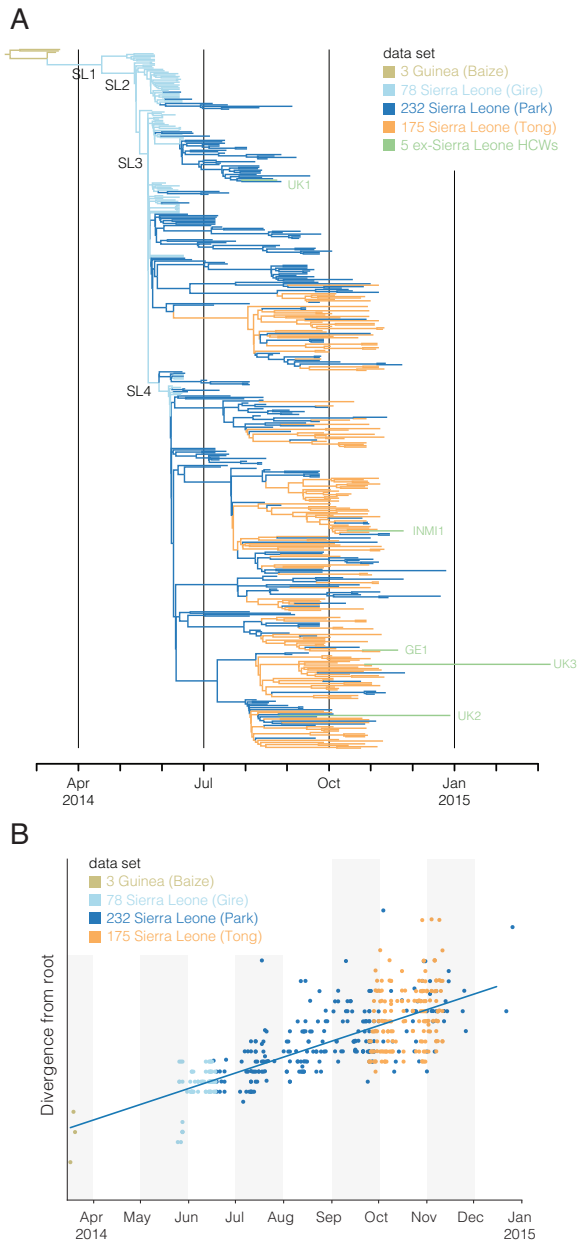


Figure B.1: Phylogenetic and temporal context of recent Tong et al. samples. (**A**) 175 recently published Ebola virus Makona samples from Sierra Leone describe lineages that fall within the genetic diversity of our current dataset (MCC tree from BEAST, as in Figure 2.1). (**B**) They span a two month period (28 September to 11 November 2014) that falls within the temporal sampling of our current data and shows a consistent evolutionary rate.

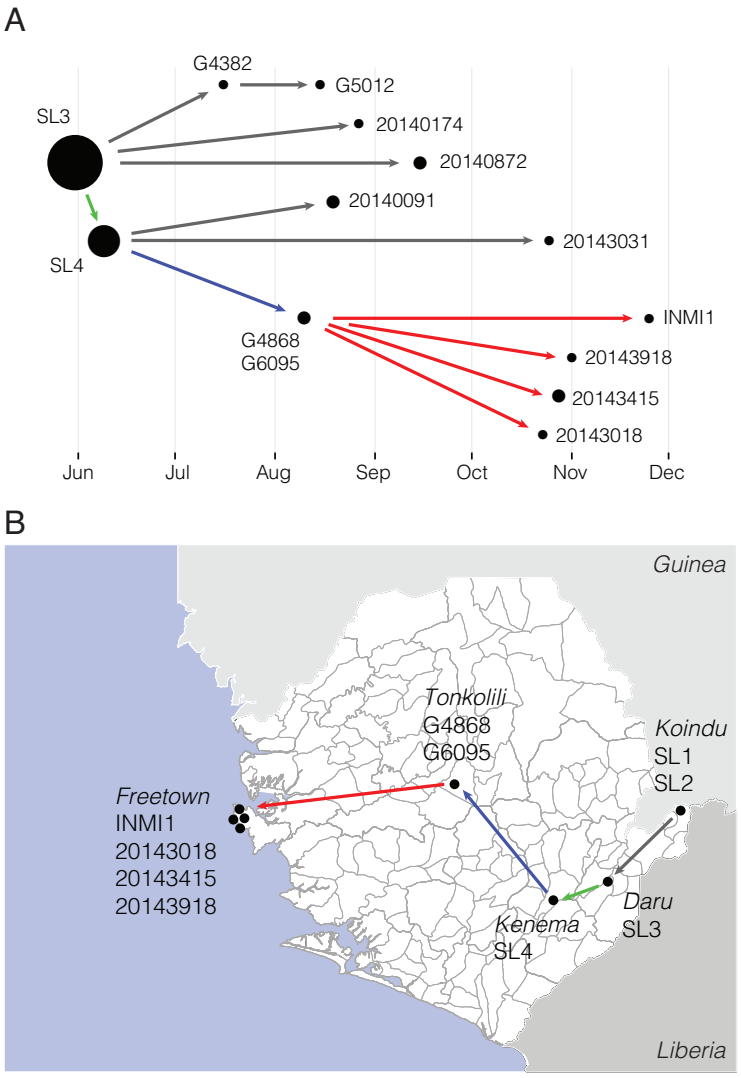


Figure B.2: Tracing historical Ebola virus Makona migrations from east to west. (**A**) Nine Ebola virus (EBOV) Makona genomes (right-hand most circles) from the Freetown area with four groups of apparently ancestral EBOV genomes (middle circles). Groups of genetically identical genomes (circles) are related to each other by simple vertical relationships (arrows). Solid circles are shown on the date of the earliest sample in the group; the circle area is proportional to the number of samples containing viruses with that genome; arrows represent a set of non-homoplasic SNPs and point from ancestral to derived alleles. Here, 'SL3' and 'SL4' do not refer to entire clades, but to the viruses that exactly match the canonical SL3 and SL4 genomes with no further mutations. (**B**) Geographic mapping of one epidemiological route that may account for four of the nine Freetown viruses shown in (A). Groups of identical viruses are shown at their first observed location.

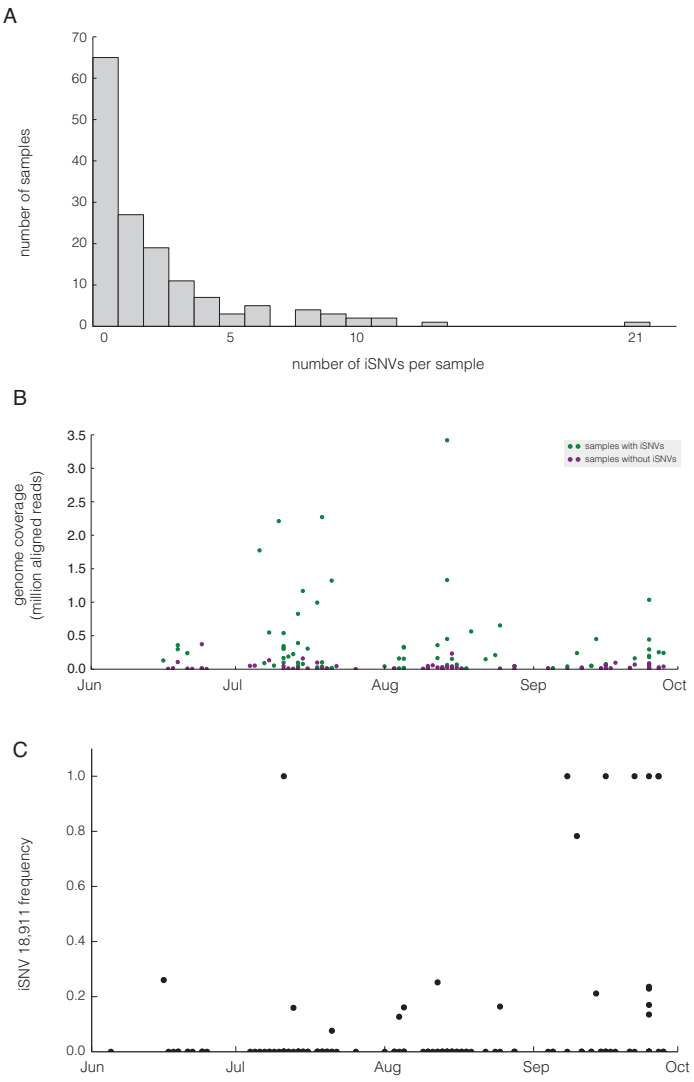


Figure B.3: Ebola virus Makona intrahost single-nucleotide variants. (**A**) Distribution of the number of iSNVs per sample. Replicate sequencing and iSNV calling was completed for 150 samples, of which 65 had no iSNV calls. Mean iSNVs per sample (including samples without iSNVs) = 2.04; mean iSNVs per sample (among samples with iSNVs) = 3.6. (**B**) Sample coverage by date shows the temporal distribution of samples containing EBOV genomes with and without iSNV calls. As expected, samples with iSNV calls have generally higher coverage. (**C**) Intermediate-frequency variants can persist over time with minimal genetic drift, as demonstrated by the iSNV at position 18,911. The existence of intermediate frequency (10%–30%) iSNVs in many different samples over time provides an argument against recurring mutations and may suggest a relatively wide transmission bottleneck between patients.

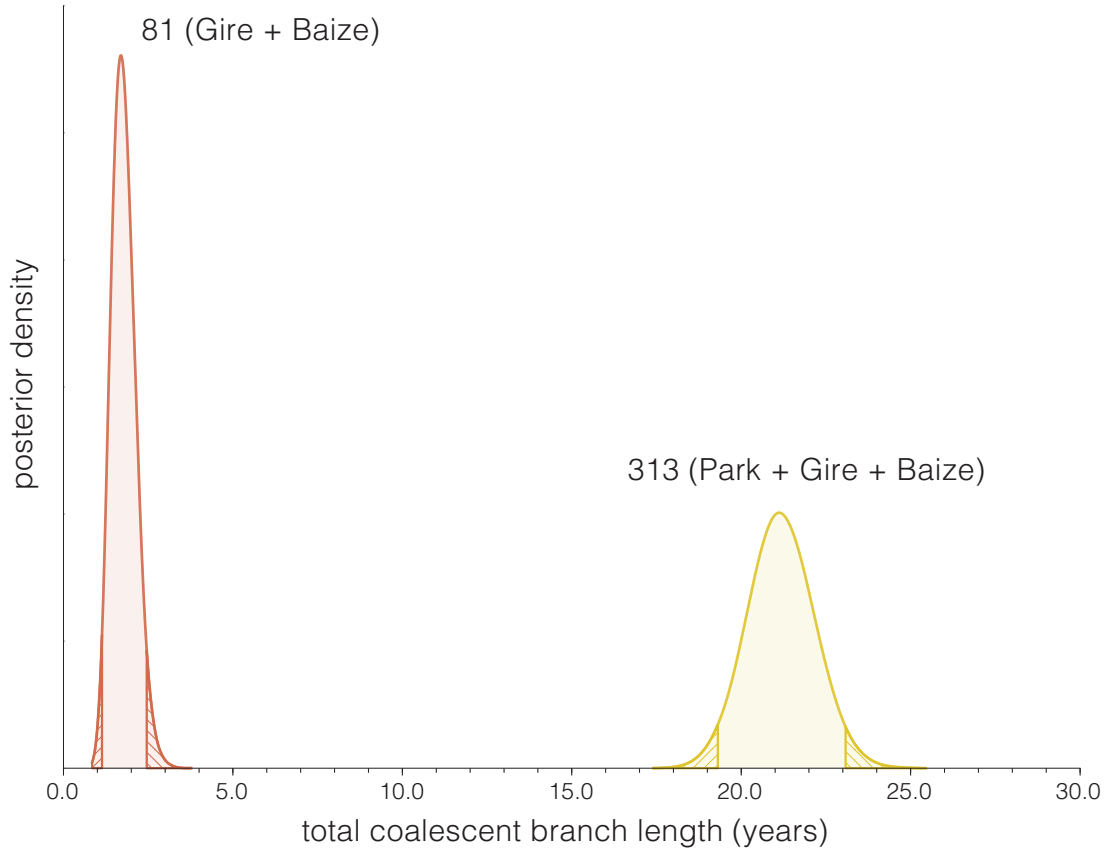


Figure B.4: Increased sampling improves evolutionary rate estimates. Rate estimates in the recent dataset (Figure 2.3A) have much tighter credible intervals due to the significantly greater amount of time (total coalescent branch length) compared to the initial outbreak.

Appendix C

Supplemental Material for Chapter 3

C.1 Supplemental Figures and Tables

Table C.1: Clinical outcome of EVD patients from Nigeria, by symptom.

Symptom		Survivor (*n*=12)	Dead (*n*=8)
Fever	Positive	11	6
	Negative	1	2
Fatigue	Positive	8	6
	Negative	4	2
Diarrhea	Positive	7	6
	Negative	5	2
Anorexia	Positive	9	2
	Negative	3	6
Vomiting	Positive	5	5
	Negative	7	3
Headache	Positive	4	2
	Negative	8	6
Bleeding	Positive	2	4
	Negative	10	4

Table C.2: Clinical profile of EVD patients from Nigeria.

Signs	Clinical Profile	Confirmed Cases (%)
Pulse Rate (*n*=12)	High (>100 beats/min)	5 (42.7)
	Normal (60-100 beats/min)	6 (50.0)
	Low (<60 beats/min)	1 (8.3)
Respiratory Rate (*n*=10)	Fast (>20 beats/min)	9 (90.0)
	Normal (12-20 beats/min)	1 (10.0)
	Slow (<12 beats/min)	0 (0.0)
Blood Pressure (*n*=11)	High (Systolic > 120)	2 (18.2)
	Normal (Systolic 90-120)	6 (54.5)
	Low (Systolic <90)	3 (27.3)
Syndromes (*n*=13)	Disseminated Intravascular Coagulation	1 (7.7)
	Gastroenteritis	9 (69.2)
	Encephalopathy	3 (23.1)

Appendix D

Supplemental Material for Chapter 4

D.1 Supplemental Figures and Tables

A

Model comparison all predictors *vs* all predictors except:	Df	2Δ ln L	Pr(>Chi-Squared)
sample site	5	21.756	<0.001***
patient gender	1	3.885	0.0487*
patient age	1	0.400	0.527
sample type	2	0.401	0.818
collection interval	3	4.749	0.191

D

Model comparison all predictors *vs* all predictors except:	Df	2Δ ln L	Pr(>Chi-Squared)
sample site	5	6.728	0.242
patient gender	1	1.642	0.200
patient age	1	0.118	0.731
sample type	2	1.121	0.571
collection interval	3	8.856	0.0313*

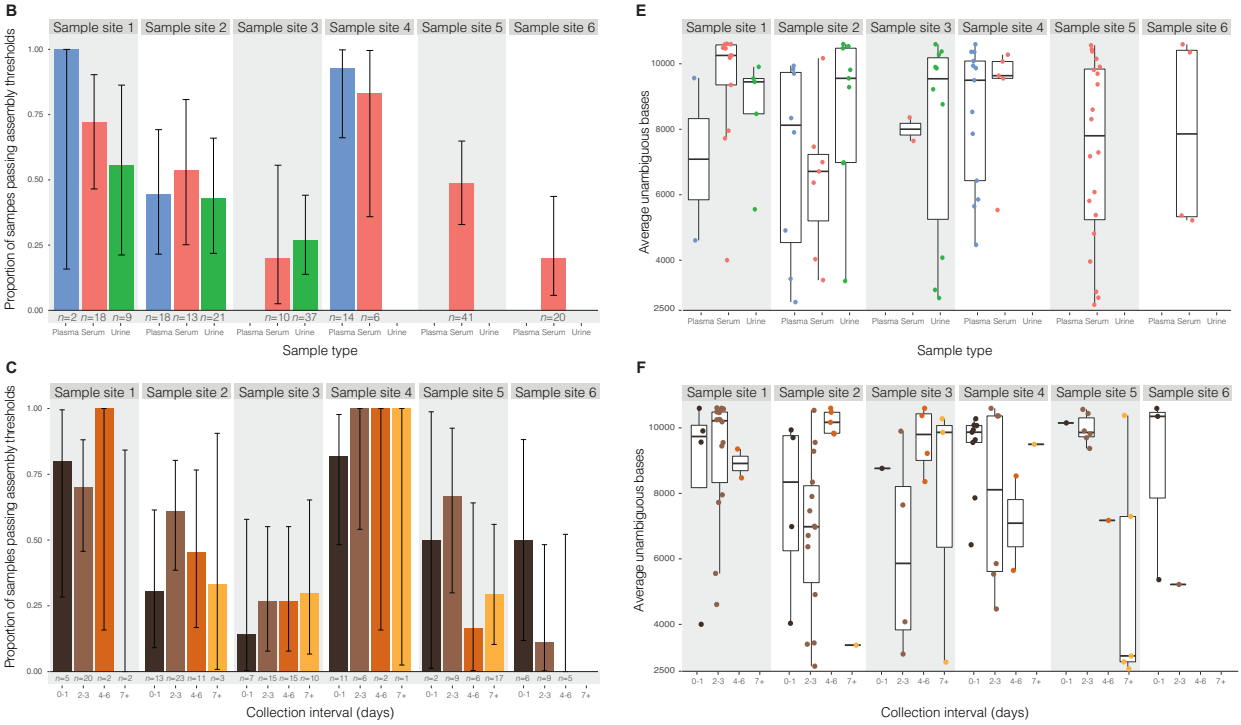


Figure D.1: Relationship between metadata and sequencing outcome. Analysis of possible predictors of sequencing outcome: the site where a sample was collected, patient gender, patient age, sample type, and collection interval. (**A**) Prediction of whether a sample will pass assembly thresholds by sequencing. Rows show results of likelihood ratio tests on each predictor by omitting the variable from a full model that contains all predictors. Sample site and patient gender improve model fit, but sample type and collection interval do not. (**B**) Proportion of samples that pass assembly thresholds by sequencing, divided by sample type, across six sample sites. (**C**) Same as B, but divided by collection interval. (**D**) Prediction of the genome fraction identified, using samples that passed assembly thresholds. Rows show results of likelihood ratio tests, as in A. Collection interval improves the model, but sample type does not. (**E**) Sequencing outcome for each sample, divided by sample type, across six sample sites. (**F**) Same as E, but divided by collection interval. Samples collected seven or more days after symptom onset produced, on average, the fewest unambiguous bases, though these observations are based on a limited number of data points. While the sample site variable accounts for differences in cohort composition, the observed effects of gender and collection interval might be due to confounders in composition that span multiple cohorts. These results illustrate the effects of variables on sequencing outcome for the samples in this study; they are not indicative of ZIKV titer more generally.

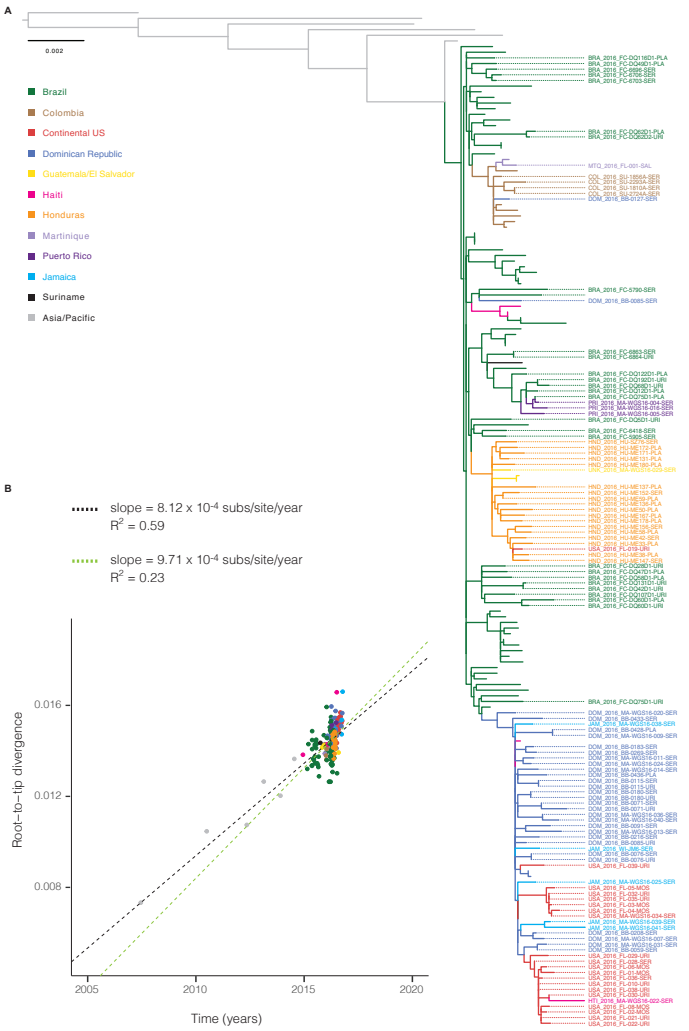


Figure D.2: Maximum likelihood tree and root-to-tip regression. (**A**) Maximum likelihood tree. Tips are coloured by sample source location. Labelled tips indicate genomes generated in this study; all other coloured tips are other publicly available genomes from the outbreak in the Americas. Grey tips are genomes from ZIKV cases in Southeast Asia and the Pacific. (**B**) Linear regression of root-to-tip divergence on dates. The substitution rate for the full tree, indicated by the slope of the black regression line, is similar to rates of Asian lineage ZIKV estimated by molecular clock analyses. The substitution rate for sequences within the Americas outbreak only, indicated by the slope of the green regression line, is similar to rates estimated by BEAST (1.15×10^{-3}; 95% CI: 9.78×10^{-4} to 1.33×10^{-3}) for this dataset.

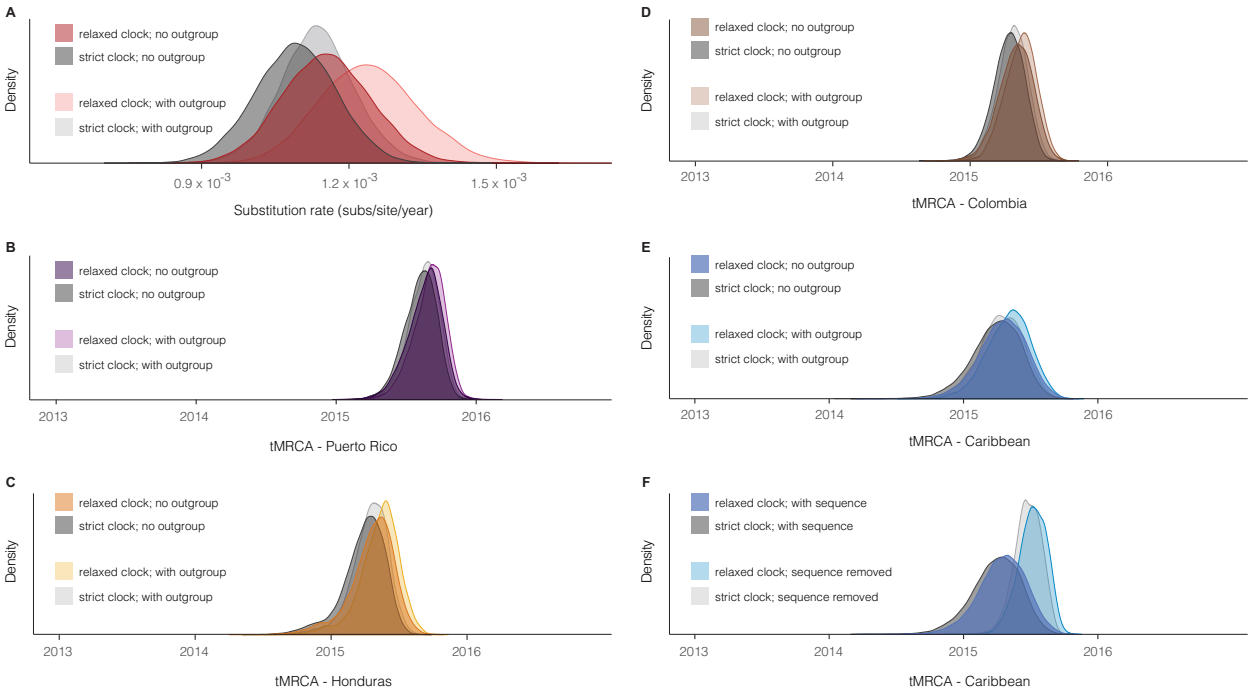


Figure D.3: Substitution rate and tMRCA distributions. (**A**) Posterior density of the substitution rate. Shown with and without the use of sequences (outgroup) from outside the Americas. (**B–E**) Posterior density of the date of the most recent common ancestor (MRCA) of sequences in four regions corresponding to those in Figure 4.2C. Shown with and without the use of outgroup sequences. The use of outgroup sequences has little effect on estimates of these dates. (**F**) Posterior density of the date of the MRCA of sequences in a clade consisting of samples from the Caribbean and continental United States. Shown with and without the sequence of DOM_2016_MA-WGS16-020-SER, a sample from the Dominican Republic that has only 3,037 unambiguous bases; this is the most ancestral sequence in the clade and its presence affects the tMRCA. In all panels, all densities are shown as observed with a relaxed clock model and with a strict clock model.

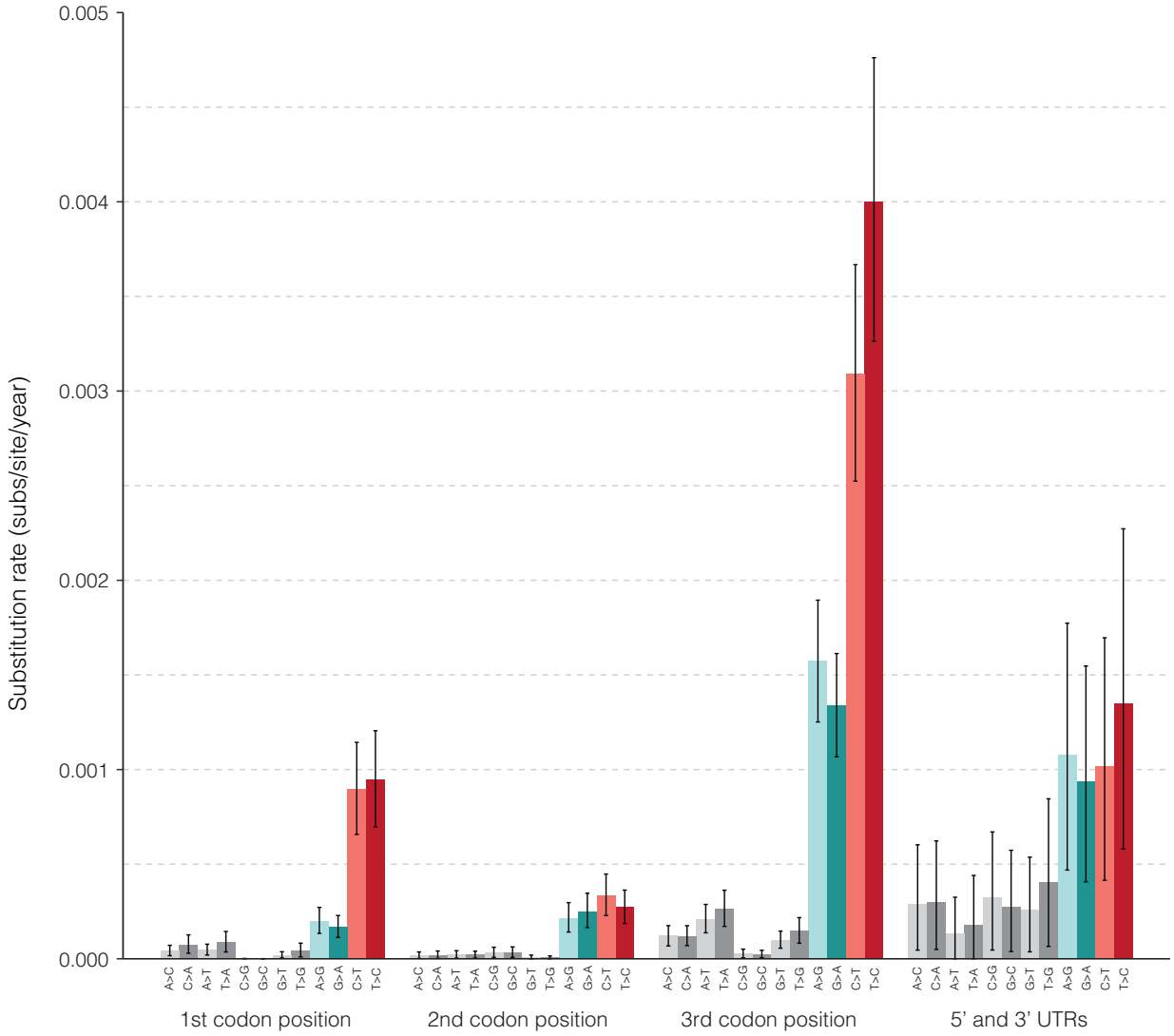


Figure D.4: Substitution rates estimated with BEAST. Substitution rates estimated in three codon positions and non-coding regions (5' and 3' UTRs). Transversions are shown in grey and transitions are coloured by transition type. Plotted values show the mean of rates calculated at each sampled Markov chain Monte Carlo (MCMC) step of a BEAST run. These calculated rates provide additional evidence for the observed high C-to-T and T-to-C transition rates shown in Figure 4.3D.

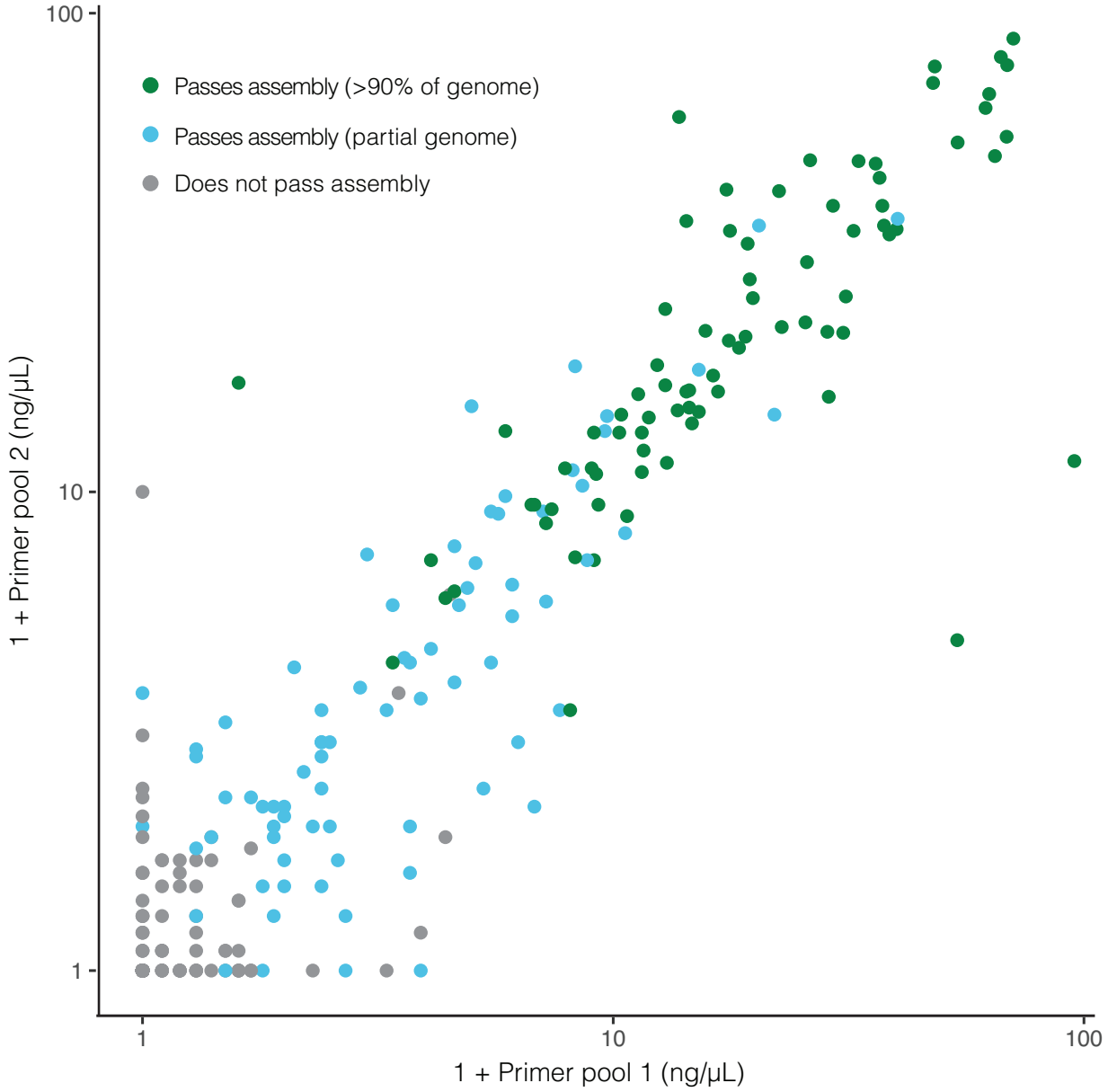


Figure D.5: cDNA concentration of amplicon primer pools predicts sequencing outcome. cDNA concentration of amplicon pools (as measured by Agilent 2200 Tapestation) is highly predictive of amplicon sequencing outcome. On each axis, 1+primer pool concentration is plotted on a log scale. Each point is a technical replicate of a sample and colours denote observed sequencing outcome of the replicate. If a replicate is predicted to be passing when at least one primer pool concentration is $\geq$0.8 ng/μ l, then sensitivity is 98.71% and specificity is 90.34%. An accurate predictor of sequencing success early in the sample processing workflow can save resources.

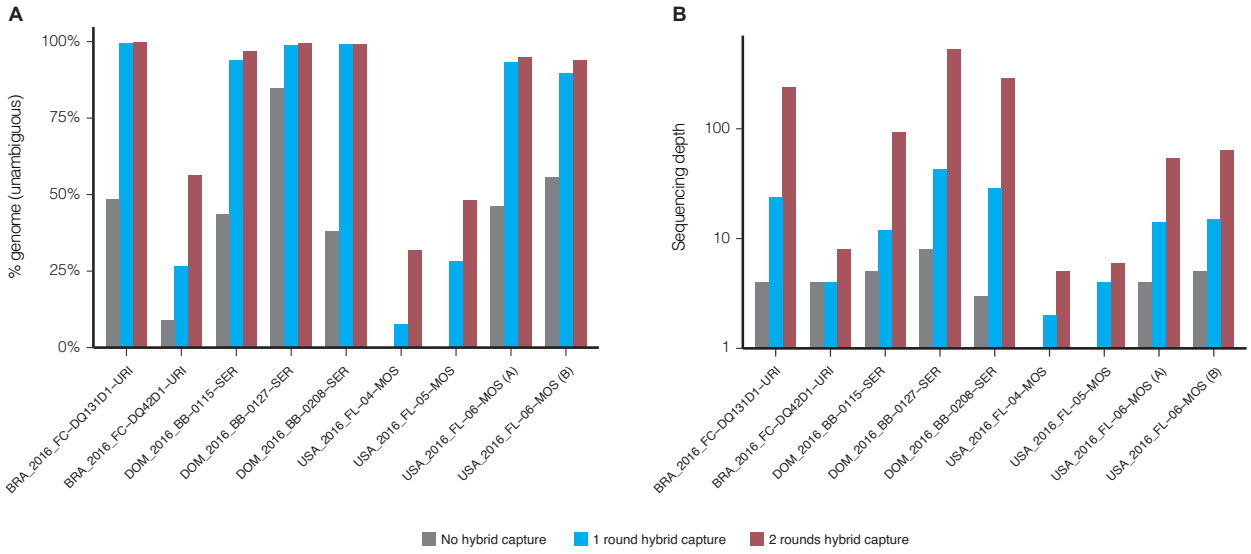


Figure D.6: Evaluating multiple rounds of Zika virus hybrid capture. Genome assembly statistics of samples before hybrid capture (grey), and after one (blue) or two (red) rounds of hybrid capture. Nine individual libraries (eight unique samples) were sequenced all three ways, had more than one million raw reads in each method, and generated at least one passing assembly. Raw reads from each method were downsampled to the same number of raw reads (8.5 million) before genomes were assembled. (**A**) Percent of the genome identified, as measured by number of unambiguous bases. (**B**) Median sequencing depth of ZIKV genomes, taken over the assembled regions.

Table D.1: Viruses other than Zika uncovered by unbiased sequencing.

A

Species	Sample	# reads from species (% of total)	% genome unambiguous
Cell fusing agent virus	USA_2016_FL-01-MOS	5662 (0.02%)	99.1%
Cell fusing agent virus	USA_2016_FL-04-MOS	1588 (0.003%)	91.1%
Cell fusing agent virus	USA_2016_FL-05-MOS	9614 (0.02%)	99.9%
Cell fusing agent virus	USA_2016_FL-06-MOS	2646 (0.007%)	82.2%
Cell fusing agent virus	USA_2016_FL-08-MOS	13608 (0.008%)	99.4%
Deformed wing virus-like	USA_2016_FL-06-MOS	6580 (0.02%)	8.34%
Dengue virus type 1	BLM_2016_MA-WGS16-006-SER	2355926 (2.6%)	99.8%
JC polyomavirus	BRA_2016_FC-DQ75D1-URI	8050 (0.20%)	99.2%
JC polyomavirus-like	USA_2016_FL-032-URI	316 (0.001%)	7.71%

B

Sample	Total contigs	Classified contigs (all)	Classified contigs (viral)	Likely novel viral contigs
USA_2016_FL-01-MOS	496	431	45	25
USA_2016_FL-02-MOS	563	463	17	14
USA_2016_FL-03-MOS	164	133	29	22
USA_2016_FL-04-MOS	679	492	25	19
USA_2016_FL-05-MOS	355	313	25	8
USA_2016_FL-06-MOS	726	635	26	14
USA_2016_FL-07-MOS	5967	5650	5	2
USA_2016_FL-08-MOS	1679	1528	39	27
All pools: unique	**9013**	**8426**	**84**	**41**

(**A**) Viral species other than Zika were found by unbiased sequencing of 38 samples. Column 3, number of reads in a sample belonging to a species as a raw count and a percent of total reads. Column 4, per cent genome assembled based on the number of unambiguous bases called. We identified cell fusing agent virus (a flavivirus) and deformed wing virus-like genomes in mosquito pools, and dengue virus type 1, JC polyomavirus, and JC polyomavirus-like genomes in clinical samples. All assemblies had $\geq$95% sequence identity to a reference sequence for the listed species, except cell fusing agent virus in USA_2016_FL-06-MOS (91%) and dengue virus type 1 in BLM_2016_MA-WGS16-006-SER (92%). The dengue virus type 1 genome showed $\geq$95% sequence identity to other available isolates of the virus. (**B**) Contigs assembled from unbiased sequencing data of eight mosquito pools. Column 2, number of contigs assembled. Column 3, number of contigs classified by BLASTN/BLASTX. Column 4, number of contigs hitting a viral species. Column 5, number of contigs hitting a viral species with <80% amino acid identity to the best hit. Each column is a subset of the previous column. Contigs in column 5 are considered to be likely to be novel. Last row lists counts, after removing duplicate contigs, for all mosquito pools combined.

Table D.2: Model selection for BEAST analyses.

A

		Skyline Relaxed	Skyline Strict	Exponential Relaxed	Exponential Strict	Constant Relaxed	Constant Strict
PS	log(marginal likelihood)	-24952	-24950	-24974	-24989	-25007	-25026
	log(Bayes factor)	74	76	53	38	20	—
SS	log(marginal likelihood)	-24957	-24954	-24976	-24990	-25010	-25030
	log(Bayes factor)	73	77	54	40	20	—

B

	Skyline Relaxed	Skyline Strict	Exponential Relaxed	Exponential Strict	Constant Relaxed	Constant Strict
Clock rate	1.15E-03 [9.78E-04, 1.33E-03]	1.09E-03 [9.32E-04, 1.25E-03]	1.06E-03 [8.38E-04, 1.29E-03]	9.42E-04 [7.42E-04, 1.14E-03]	1.41E-03 [1.15E-03, 1.69E-03]	1.18E-03 [9.97E-04, 1.36E-03]
tMRCA: all	2014.129 [2013.621, 2014.552]	2013.981 [2013.531, 2014.417]	2013.498 [2012.772, 2014.175]	2013.401 [2012.724, 2014.028]	2013.752 [2012.897, 2014.405]	2013.806 [2013.349, 2014.241]
tMRCA: Puerto Rico	2015.632 [2015.376, 2015.849]	2015.600 [2015.369, 2015.816]	2015.599 [2015.314, 2015.900]	2015.530 [2015.231, 2015.832]	2015.796 [2015.533, 2016.039]	2015.714 [2015.491, 2015.951]
tMRCA: Honduras	2015.300 [2014.928, 2015.594]	2015.241 [2014.888, 2015.512]	2015.197 [2014.850, 2015.524]	2015.066 [2014.684, 2015.392]	2015.527 [2015.206, 2015.834]	2015.334 [2015.049, 2015.599]
tMRCA: Colombia	2015.333 [2015.088, 2015.567]	2015.283 [2015.060, 2015.496]	2015.246 [2014.989, 2015.472]	2015.153 [2014.873, 2015.398]	2015.411 [2015.201, 2015.636]	2015.306 [2015.096, 2015.503]
tMRCA: Caribbean	2015.289 [2014.933, 2015.628]	2015.242 [2014.876, 2015.578]	2015.140 [2014.798, 2015.465]	2015.007 [2014.623, 2015.373]	2015.412 [2015.073, 2015.754]	2015.278 [2014.952, 2015.605]

(**A**) Marginal likelihoods calculated with path-sampling (PS) and stepping-stone sampling (SS) for combinations of three coalescent tree priors (constant size population, exponential growth population, and Skyline) and two clock models (strict clock and uncorrelated relaxed clock with log-normal distribution). The Bayes factor is calculated against the baseline model, a constant size tree prior and strict clock. (**B**) Mean estimates and 95% credible intervals across evaluated models for the clock rate, date of tree root, and tMRCAs of the four regions shown in Figure 4.2C. Under a Skyline tree prior, the use of strict and relaxed clock models yields similar estimates.

Table D.3: Within-sample variant validation.

A

Method	% unvalidated by other method
Amplicon sequencing	87.3% $n = 126$
Hybrid capture	85.8% $n = 113$
Hybrid capture, verified	25.0% $n = 20$

B

Method	% unvalidated in replicate	
	all variants	variants passing strand bias filter
Amplicon sequencing	92.7% $n = 304$	66.7% $n = 3$
Hybrid capture	74.5% $n = 98$	0.00% $n = 8$

(**A**) For each method (amplicon sequencing or hybrid capture), fraction of identified variants ($\geq$1%) not identified at $\geq$1% by the other method (that is, unvalidated). 'Verified' hybrid capture variants are those passing strand bias and frequency filters, as described in Methods, Section 4.5. (**B**) For each method, the fraction of identified variants unvalidated in a second library. To pass the strand bias filter, a variant must meet filter criteria in both replicates.

Appendix E

Supplemental Material for Chapter 5

E.1 Supplemental Figures and Tables

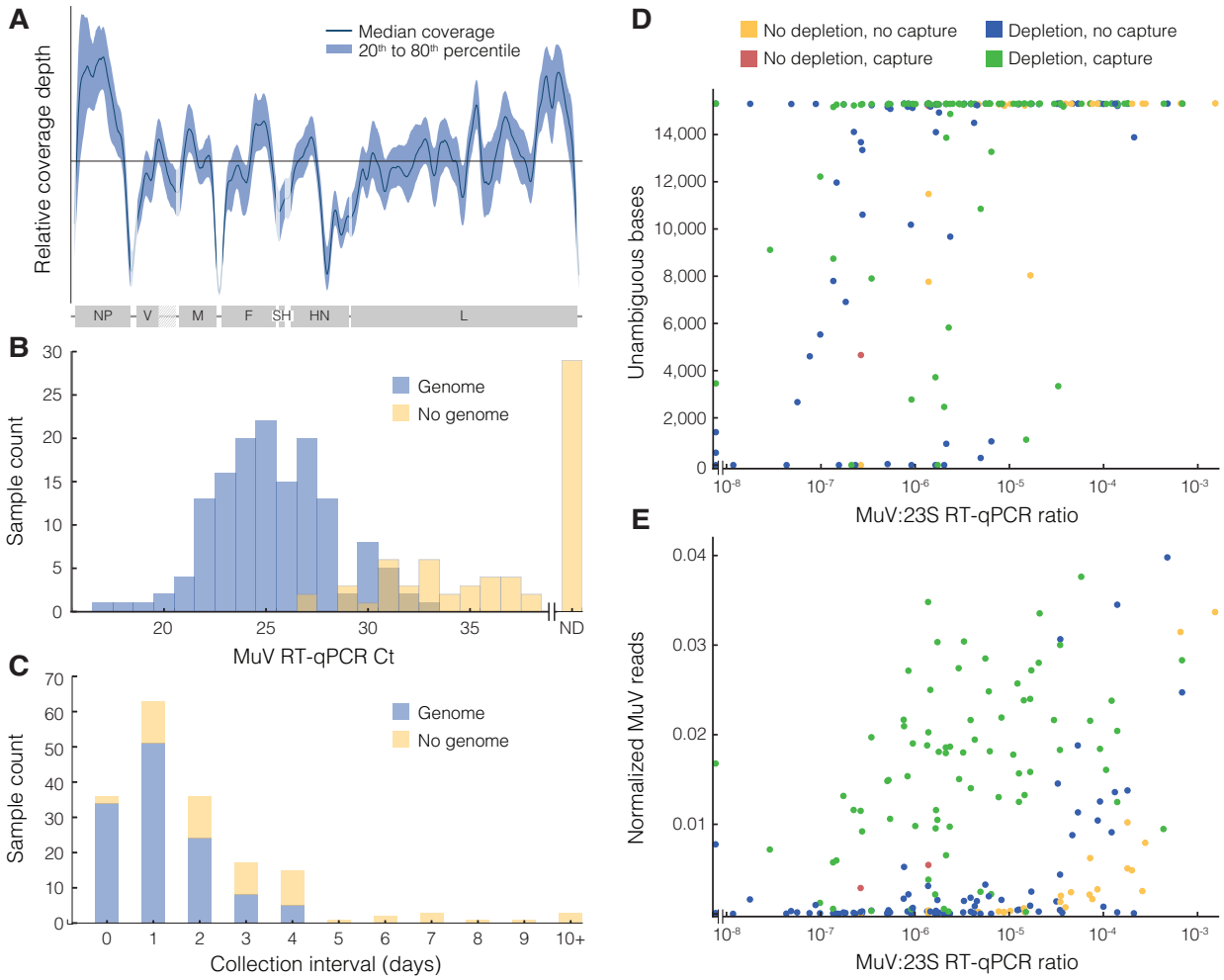
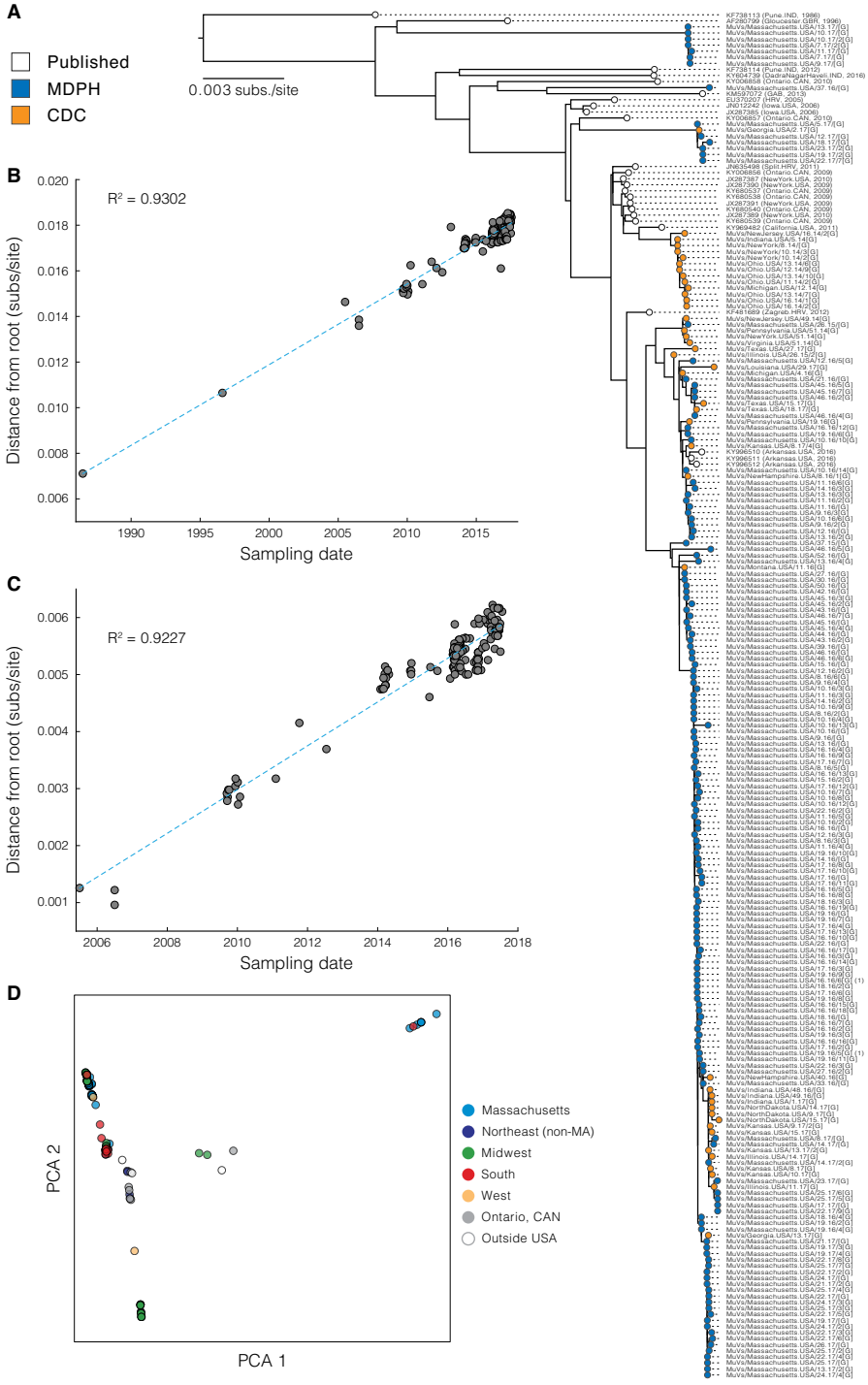


Figure E.1: Sequencing results and predictors of outcome. (**A**) Relative sequencing depth of coverage aggregated across 200 MuV genomes. (**B**) Distribution of MuV RT-qPCR cycle threshold (Ct) value, taken at sample source (MDPH or CDC), for all samples prepared with both depletion and capture (see Methods, Section 5.5). Samples that produced genomes (blue) passed the thresholds described in Methods. MuV RT-qPCR serves as a predictor of sequencing outcome. (**C**) Distribution of collection interval (days between symptom onset and sample collection) for all samples prepared with both depletion and capture. Samples that produced genomes (blue) passed the thresholds described in Methods. Samples taken more than 4 days after symptom onset did not produce genomes in this study. (**D**) Number of unambiguous bases in a genome assembly by MuV:23S ratio (MuV copies by MuV RT-qPCR divided by 23S copies by 23S RT-qPCR; see Methods). Each point is a replicate, colored by sequencing preparation method. (**E**) Unique MuV reads for a sample, divided by raw sequencing depth, by MuV:23S ratio. Points are as in panel (D). Nine points with >0.04 normalized MuV reads are not shown. In both (D) and (E), four points with a ratio $<10^{-8}$ are shown at 0.

Figure E.2: Maximum likelihood tree, root-to-tip regression, and principal components analysis. (**A**) Maximum likelihood tree of the 225 MuV genotype G genomes used in this study. Tips are colored by sample source (MDPH or CDC); previously-published genomes are indicated by unfilled circles. (**B**) Root-to-tip regression of genomes shown in (A), rooted on GenBank accession: KF738113 (Pune.IND, 1986). (**C**) Root-to-tip regression of genomes in the clade containing the two USA 2006 sequences (USA_2006; see Figure 5.1A), as well as their descendants. (**D**) Principal components analysis of genetic variants from the genomes in (A). Each point is a genome colored by its geographic location.

Figure E.2: Continued



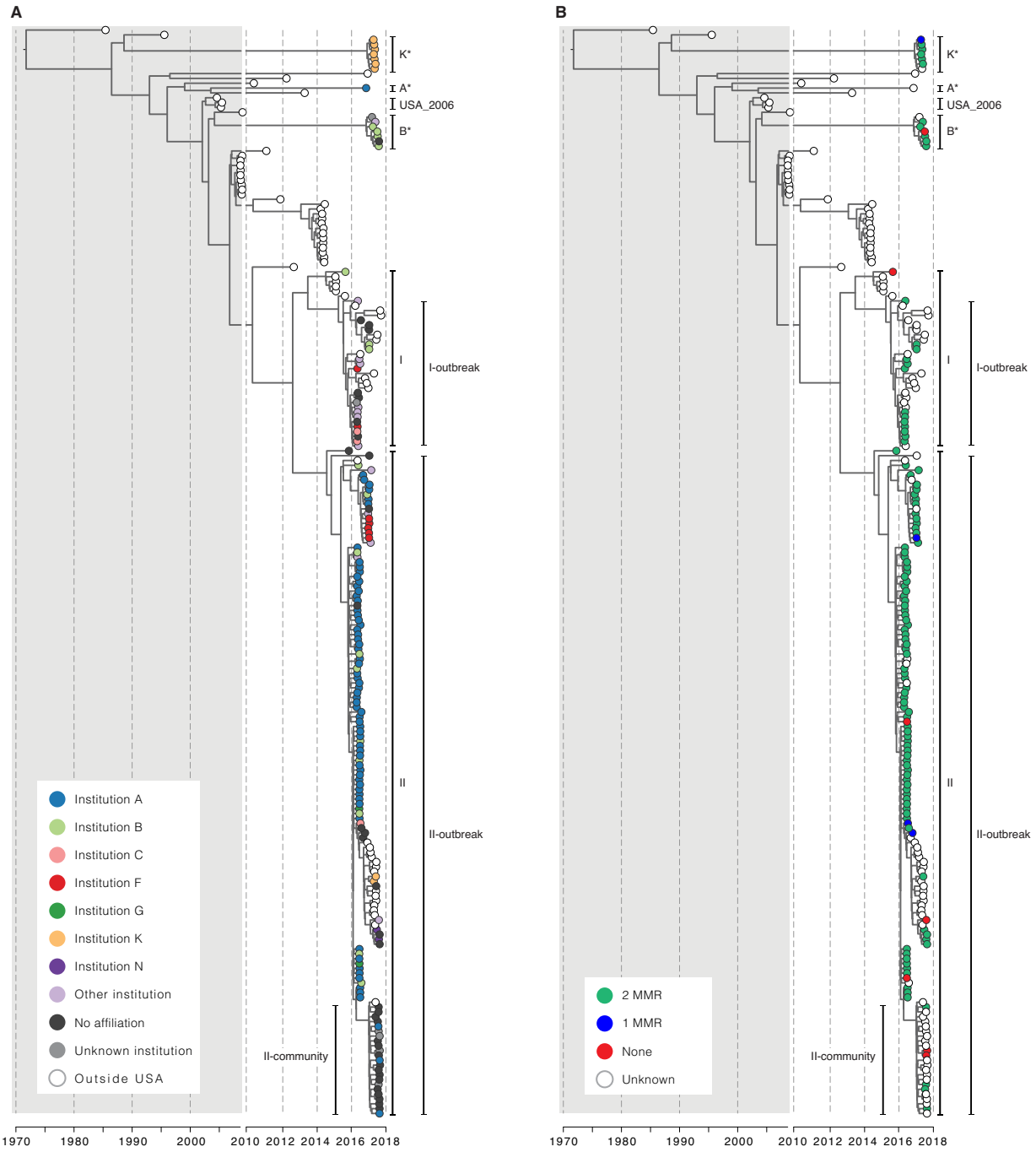


Figure E.3: Phylogenetic trees colored by institution and vaccination status. Maximum clade credibility tree of the 225 MuV genotype G genomes used in this study, colored by (**A**) academic institution and (**B**) MMR vaccination status. Clades are labeled as in Figure 5.1A.

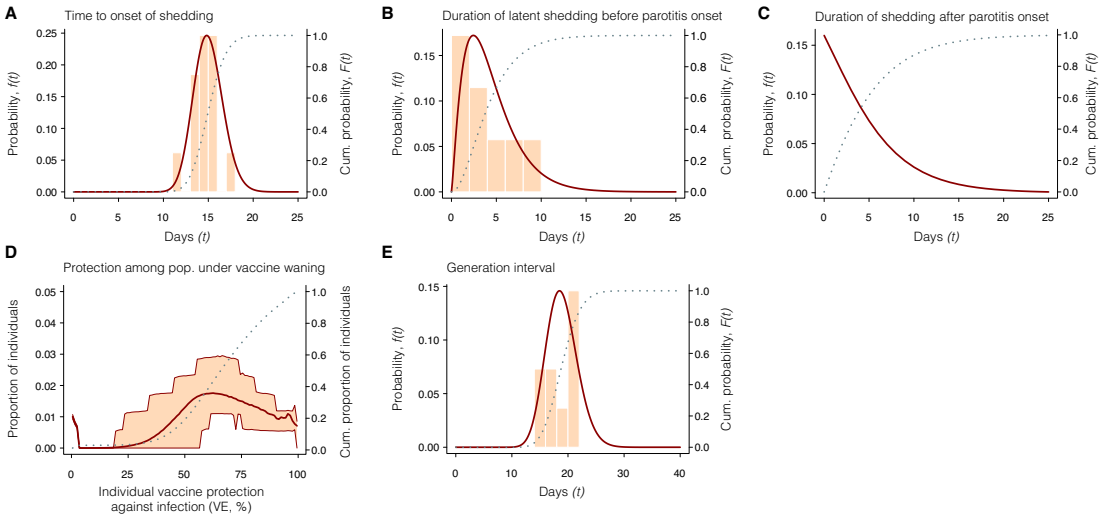


Figure E.4: Parameters used in epidemiological models. We illustrate fitted distributions of parameters of the modeled natural history of mumps infection. We use published data (see Methods) to fit (**A**) the distribution of incubation period (the time of mumps virus exposure to the onset of shedding), (**B**) the distribution of the period of latent shedding, (**C**) the distribution of the duration of shedding after parotitis onset, and (**D**) the distribution of vaccine protection within a university protection. (**E**) Distribution of the generation interval, fit using 10 samples in our dataset with known exposure sources.

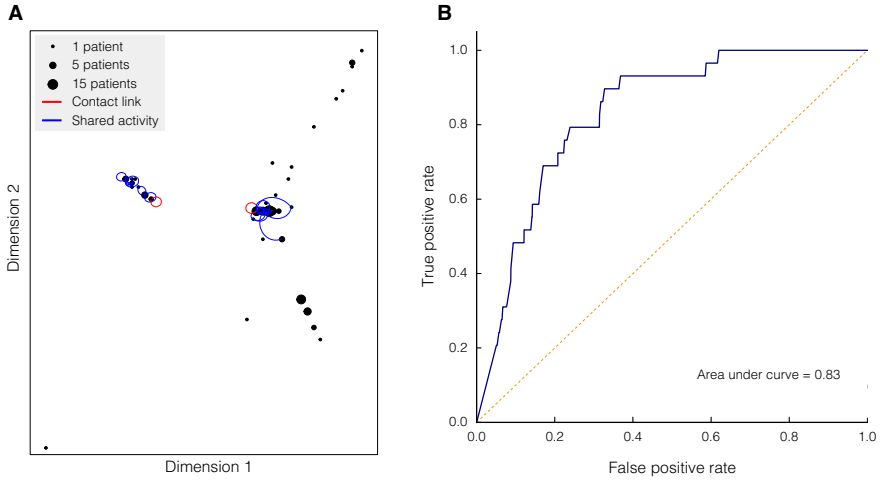


Figure E.5: Connection between epidemiological and genetic data. (**A**) Multidimensional scaling applied to samples in clade II-outbreak (see Figure 5.1A). Each point is a MuV genome and pairwise dissimilarities are based on Hamming distance. Genomes with known epidemiological links are connected with a colored line. (**B**) Receiver operating characteristic (ROC) curve for samples within clade II-outbreak using pairwise genetic distance (calculated as in (A)) as a predictor of epidemiological linkage.

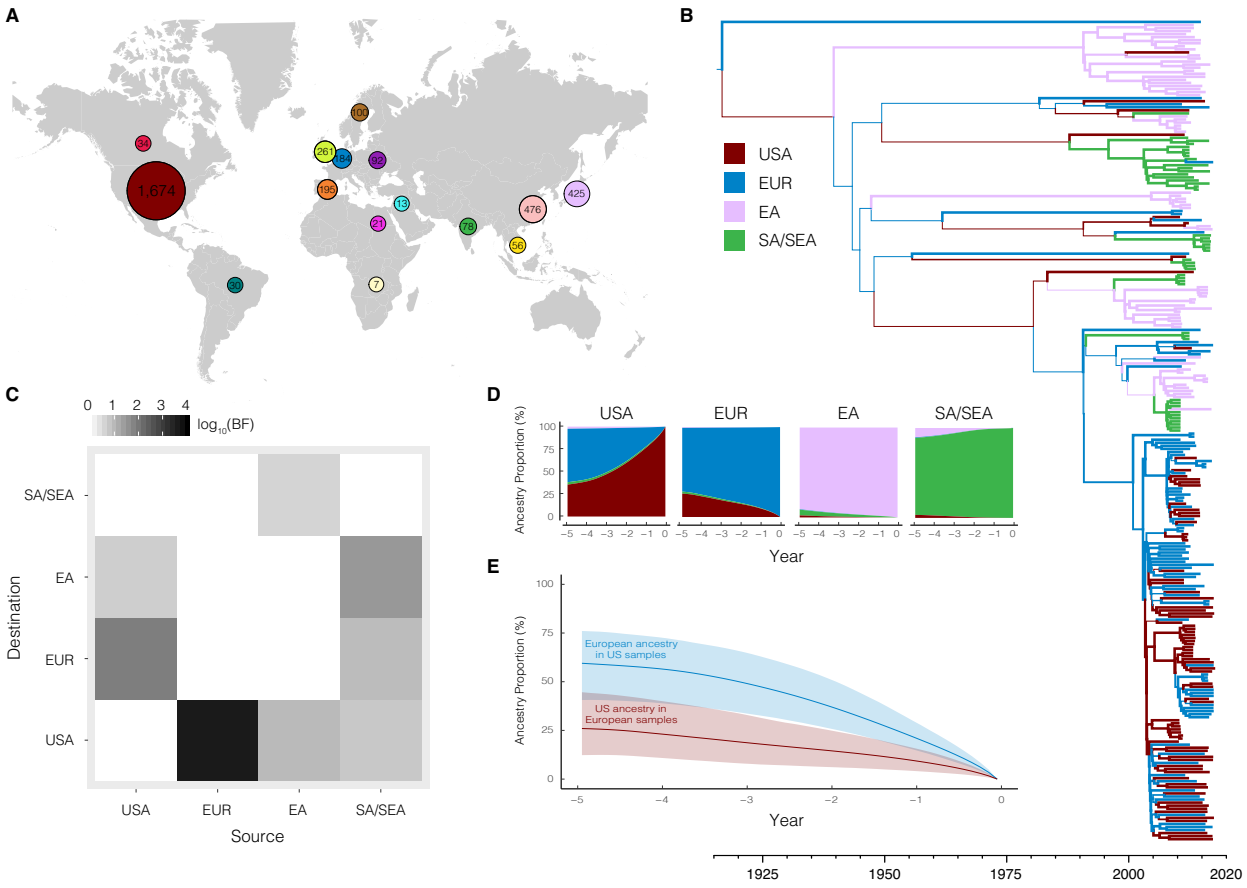


Figure E.6: Additional analyses of global MuV spread using SH gene sequences. (**A**) World map indicating number of SH sequences in our dataset from each of 15 regions. (**B**) Tree with the highest clade credibility across all trees generated on resampled input from four world regions: South Asia/Southeast Asia (SA/SEA), East Asia (EA), Europe (EUR), and United States (USA) (see Methods for details regarding geographic and temporal resampling of sequences). Branch line thickness corresponds to posterior support for ancestry (indicated by branch color). (**C**) Migration between the 4 regions shown in (A). Shading of each migration route indicates its statistical support (quantified with Bayes factors) in explaining the diffusion of MuV. (**D**) Average proportion of geographic ancestry of samples in each of the four world regions (labeled) from each of the four regions (colored), going back five years from sample collection. Colors are as in (B). (**E**) Average proportion of EUR in geographic ancestry of USA samples, and vice-versa. Shaded regions are pointwise percentile bands (2.5% to 97.5%) across 100 resamplings of the input sequences.

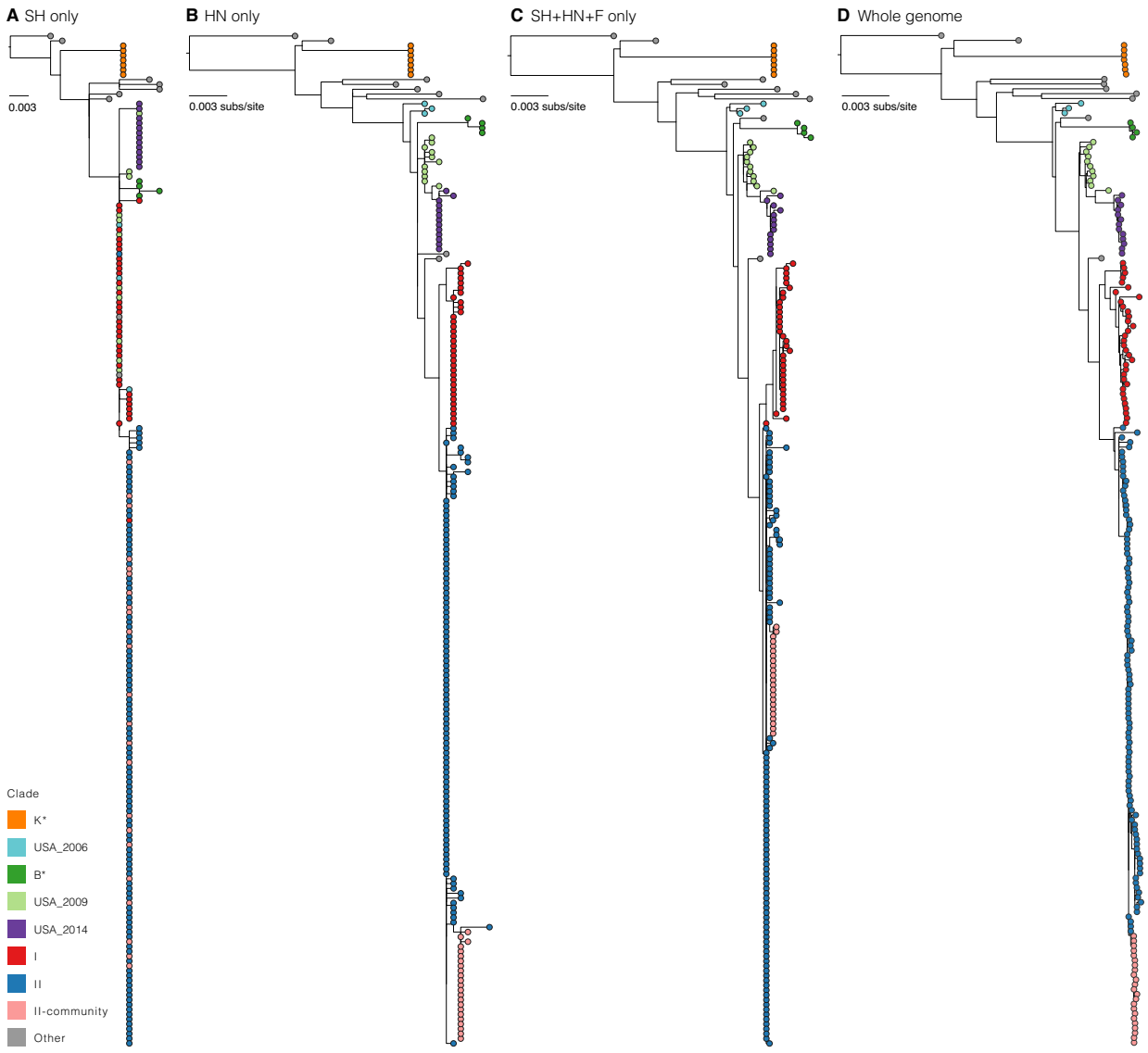


Figure E.7: Trees produced with single- and multi-gene sequences. Maximum likelihood trees using (**A**) the SH gene only, (**B**) the HN CDS only, (**C**) a concatenation of the HN CDS, the fusion protein (F) CDS, and the SH gene, and (**D**) the complete MuV genome. In all panels, tips are colored by clades as defined in Figure 5.1A and Table E.4. The HN protein sequence does a significantly better job at capturing the epidemiologically-relevant clades than the SH gene, and the tree created from SH+HN+F (nearly 25% of the genome) closely resembles the tree created from whole genome sequences.

Table E.1: Sample and genome counts.

Sample source	Collection dates	PCR result	Samples	Genomes
CDC	2014–2015	+	59	43
MDPH	2014–2015	+	6	2
MDPH	2016–2017	+	194	158
MDPH	2016–2017	–	29	0

CDC = Centers for Disease Control and Prevention; MDPH = Massachusetts Department of Public Health. Genomes meet the thresholds described in Methods.

Table E.2: Viruses other than MuV identified in PCR-negative samples.

Pathogen	Sample	Taxon-filtered reads	Contigs	Unambiguous bases	Median coverage	Genome coverage
Human Parainfluenza Virus 3	MA/24.16/NEG-1	2,560	6	13,859	8	89.6%
Human Parainfluenza Virus 2	MA/39.16/NEG-1	294	5	210	0	1.3%
Human Coronavirus OC43	MA/50.16/NEG-1	46,136	12	1,976	1	6.5%
Influenzavirus B	MA/10.16/NEG-1	54*	0	–	–	–

Influenzavirus B is the only multipartite virus listed, and we identify 51 reads mapping to six of the eight segments: in order, 2, 6, 0, 0, 8, 10, 4, 21 reads to each segment.

Table E.3: Sample metadata.

A

		All cases (n=377)	Study cases (n=198)
Gender	Male	49.07%	49.50%
	Female	50.93%	50.50%
Age	<5 years	1.86%	0.00%
	5–9 years	1.06%	0.00%
	10–14 years	27.00%	0.51%
	15–19 years	21.49%	23.23%
	20–24 years	42.97%	48.48%
	25–29 years	12.20%	11.62%
	30+ years	20.16%	16.16%
Vaccination status	2+ MMR	64.97%	64.14%
	1 MMR	4.55%	3.03%
	Unvaccinated	4.55%	5.56%
	Unknown	25.94%	27.27%

B

		Negatives (n=521)
Vaccination status	2+ MMR	47.60%
	1 MMR	7.87%
	Unvaccinated	3.84%
	Unknown	40.69%
Time since vaccination	<1 year	2.69%
	1–4 years	9.21%
	5–9 years	7.10%
	10–14 years	11.71%
	15–19 years	18.43%
	20–24 years	4.41%
	25+ years	1.92%
	Unvaccinated	3.84%
	Unknown	40.69%
Time to collection	0 days	18.04%
	1 day	30.71%
	2 days	19.96%
	3 days	9.79%
	4 days	6.33%
	5+ days	6.72%
	Unknown	8.45%

(**A**) Demographic information of all MuV cases in MA between 2016-01-01 and 2017-06-30, and the subset of these included in this study. (**B**) Relevant metadata for all samples PCR-negative for MuV collected during the same time period, 29 of which we attempted to sequence.

Table E.4: Model selection and tMRCA estimates across models.

A

		Skygrid Relaxed	Skygrid Strict	Exponential Relaxed	Exponential Strict	Constant Relaxed	Constant Strict
PS	log(marginal likelihood)	-33344	-33349	-33348	-33364	-33378	-33400
	log(Bayes factor)	57	52	52	36	22	—
SS	log(marginal likelihood)	-33332	-33352	-33350	-33367	-33380	-33404
	log(Bayes factor)	73	52	55	37	24	—

B

	Skygrid Relaxed	Skygrid Strict	Exponential Relaxed	Exponential Strict	Constant Relaxed	Constant Strict
Clock rate (x 10^{-4})	4.76	4.02	4.99	4.04	5.75	4.09
	[3.97, 5.61]	[3.60, 4.44]	[4.14, 5.85]	[3.66, 4.47]	[4.80, 6.72]	[3.68, 4.50]
tMRCA: all	1973.02	1968.844	1973.784	1969.005	1976.862	1969.522
	[1962.05, 1982.491]	[1964.95, 1972.567]	[1963.766, 1982.305]	[1965.388, 1972.809]	[1966.405, 1986.424]	[1966.006, 1973.212]
tMRCA: USA-4	2012.667	2011.977	2012.705	2012.06	2012.815	2012.045
	[2011.436, 2013.688]	[2011.058, 2012.701]	[2011.704, 2013.697]	[2011.282, 2012.752]	[2011.814, 2013.776]	[2011.382, 2012.738]
tMRCA: I	2013.535	2013.021	2013.558	2012.995	2013.621	2012.962
	[2012.816, 2014.239]	[2012.331, 2013.646]	[2012.758, 2014.279]	[2012.325, 2013.584]	[2012.848, 2014.332]	[2012.344, 2013.586]
tMRCA: I-outbreak	2015.568	2015.452	2015.518	2015.386	2015.457	2015.288
	[2015.343, 2015.787]	[2015.208, 2015.667]	[2015.226, 2015.771]	[2015.1, 2015.645]	[2015.151, 2015.736]	[2014.996, 2015.586]
tMRCA: II	2014.621	2014.147	2014.627	2014.175	2014.571	2014.089
	[2013.928, 2015.291]	[2013.553, 2014.661]	[2014.008, 2015.216]	[2013.655, 2014.711]	[2013.91, 2015.221]	[2013.531, 2014.591]
tMRCA: II-outbreak	2014.88	2014.352	2014.864	2014.375	2014.833	2014.282
	[2014.246, 2015.477]	[2013.808, 2014.887]	[2014.274, 2015.394]	[2013.869, 2014.916]	[2014.221, 2015.403]	[2013.752, 2014.771]
tMRCA: II-community	2017.043	2016.999	2017.033	2016.995	2016.942	2016.929
	[2016.871, 2017.184]	[2016.817, 2017.149]	[2016.857, 2017.186]	[2016.828, 2017.164]	[2016.711, 2017.138]	[2016.729, 2017.11]
tMRCA: K*	2016.942	2016.915	2016.933	2016.914	2016.875	2016.882
	[2016.778, 2017.083]	[2016.738, 2017.075]	[2016.753, 2017.083]	[2016.715, 2017.064]	[2016.614, 2017.066]	[2016.667, 2017.067]
tMRCA: USA_2006	2003.902	2003.621	2004.208	2003.722	2004.676	2003.972
	[2002.601, 2005.032]	[2002.747, 2004.37]	[2003.121, 2005.151]	[2002.909, 2004.448]	[2003.777, 2005.472]	[2003.246, 2004.63]
tMRCA: B*	2016.89	2016.841	2016.885	2016.833	2016.823	2016.776
	[2016.715, 2017.03]	[2016.624, 2017.019]	[2016.694, 2017.03]	[2016.604, 2017.022]	[2016.547, 2017.028]	[2016.503, 2017.006]

(**A**) Marginal likelihoods estimated in six models: combinations of three coalescent tree priors (constant size population, exponential growth population, and Skygrid) and two clock models (strict clock and uncorrelated relaxed clock with log-normal distribution). Estimates are with path-sampling (PS) and stepping-stone sampling (SS). The Bayes factors are calculated against the model with constant size population and a strict clock. (**B**) Mean estimates of clock rate, date of tree root, and tMRCAs of the clades shown in Figure 5.1A (excluding clade A*, which consists of one sample). USA-4 corresponds to 'Clades I and II' in Figure 5.1A. Below each mean estimate is the 95% credible interval.